The Synthesizing Capacity of Metabolic Networks

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach Biophysik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Herr MSc. Thomas Handorf
geboren am 28.05.1977 in Berlin

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Dr. h.c. Christoph Markschies

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Dr. Christian Limberg

Gutachter:

1. Prof. Dr. Hermann-Georg Holzhütter, Charité Berlin
2. Prof. Dr. Stefan Schuster, Universität Jena
3. Prof. Dr. Daniel Kahn, Université Lyon 1

eingereicht am: 19.11.2007
Tag der mündlichen Prüfung: 23.07.2008

Abstract

In this work, the concept of scopes is introduced and applied to large scale metabolic networks. The scopes represent functional measures, describing the synthesizing capacity of a metabolic network if supplied with a predefined set of resources. For a given set of initial metabolites, the seed, all possible products are determined using the stoichiometric information of the network. Specifically, the organism independent KEGG reference network is analyzed.

The first part of this work describes possible applications of the scopes, including the determination of the synthesizing capacities of different compounds and sets of compounds, the study of the effect of cofactors on the capacities of metabolic networks or the identification of possible nutrient sets required for the maintenance of a cell.

In the second part, the scopes of different seed compounds are systematically analyzed and put in relation to one another. A hierarchy is generated representing the inclusion relations of the scopes. Interestingly, this hierarchy reflects the chemical composition, i.e. the chemical elements or chemical groups of the contained compounds. Scopes containing frequently used chemical elements or groups are represented by high degree nodes in this hierarchy. A subhierarchy of these characteristic scopes is described and brought in relation to the autotrophy of the network.

In the third part, the effect of modifications in the topology of metabolic networks is analyzed. It turns out that the scopes are generally robust against the deletion of single and even multiple reactions. It is further investigated, how the scope hierarchies depend on the number of reactions in the network. As a result, the KEGG network appears to optimized in order to provide a sufficient number of chemical transformations while keeping the number of reactions, and hence of the corresponding enzymes, small.

Also, the influence of limitations in the metabolic knowledge on the results is discussed and possibilities for improvements are indicated. The performed analyses reveal evolutionary objectives behind the construction of metabolic networks. In particular, hypotheses about design, autotrophy or robustness of metabolic networks can be inferred.

Keywords:
metabolic network, structural analysis, synthesizing capacity, metabolic hierarchy

Zusammenfassung

In dieser Arbeit wird das Konzept der Scopes und auf großskalige metabolische Netzwerke angewendet. Mit Scopes ist es möglich, funktionelle Aussagen über solche Netze zu machen. Sie beschreiben die Synthesekapazität eines Netzwerkes, wenn dieses mit bestimmten Ausgangsstoffen versorgt wird. Dabei werden für eine bestimmte Kombination von Ausgangsstoffen alle durch das Netzwerk synthetisierbaren Stoffe berechnet. In dieser Arbeit wird insbesondere das Referenznetzwerk der KEGG-Datenbank untersucht, welches Reaktionen unabhängig von ihrem Vorkommen in unterschiedlichen Organismen enthält.

Im ersten Teil werden die Synthesekapazitäten systematisch für alle Einzelstoffe und für einige Stoffkombinationen errechnet und untersucht. Desweiteren wird der Effekt von Kofaktoren analysiert. Durch eine Inversion des Konzeptes ist es möglich, Kombinationen von Ausgangsstoffen zu finden, aus denen bestimmte wichtige Metabolite der Zelle produziert werden können. Somit kann der Nährstoffbedarf einer Zelle abgeschätzt werden.

Im zweiten Teil werden die Scopes selbst analysiert und zueinander in Relation gesetzt. Es wird eine Hierarchie der Scopes, basierend auf Inklusionen zwischen diesen, erstellt. Diese Hierarchie kann mit der chemischen Komposition der enthaltenen Stoffe, also mit deren chemischen Bausteinen, den Elementen oder Gruppen, in Verbindung gebracht werden. Dabei erhalten Scopes mit sehr häufigen Bausteinkombinationen eine hervorgehobene Rolle in der Hierarchie. Diese charakteristischen Scopes zeigen eine Unterhierarchie die mit der Autotrophie des Netzwerkes in Zusammenhang gebracht werden kann.

Der dritte Teil beschäftigt sich mit möglichen Änderungen in der Topologie des Netzwerkes und deren Auswirkungen auf die Scopes. Es stellt sich heraus, dass die Synthesekapazitäten sich im allgemeinen sehr robust gegenüber solchen Veränderungen verhalten. Ähnlich verhält es sich auch mit den Scope-Hierarchien. Die Anzahl der Reaktionen im KEGG-Netzwerk ist aber offensichtlich trotzdem dahingehend optimiert, dass eine zu große Zahl von Reaktionen und damit an alternativen Routen vermieden wird.

Außerdem wurde die Auswirkung der Unvollständigkeit des derzeitigen biochemischen Wissens auf die in dieser Arbeit präsentierten Ergebnisse diskutiert. Die Methodik ist im übrigen auch geeignet um Lücken in diesem Wissen aufzuspühren und dadurch die Kenntnisse über den Metabolismus zu erweitern. Die getätigten Analysen zeigen evolutionäre Ziele hinter der

Konstruktion metabolischer Netzwerke auf. Insbesondere konnten Hypothesen über das Design, die Autotrophy und Robustheit des Metabolismus abgeleitet werden.

Schlagwörter:
Metabolisches Netzwerk, Strukturelle Analyse, Synthesekapazität, Metabolische Hierarchie

In memory of my mentor Prof. Dr. Reinhart Heinrich.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cellular organisms synthesize their basic components and energy carriers by taking up resources from the environment. The processes that transform these resources into the desired end products are performed by a network of linked chemical reactions, commonly known as the metabolism. The metabolisms of different species may vary drastically, but they also share common principles concerning for example the topology of the network, the types of reactions and important cofactors. In many organisms, such as E.coli or Homo sapiens, the number of participating chemical reactions easily exceeds 1000.

Most reactions occurring in organisms are of enzymatic nature. One reason for this is that non catalyzed reactions, transforming relatively complex chemical compounds into one another, are rather unlikely, resulting in a very slow reaction kinetics. In physiological time scales one can therefore assume that such reactions do not occur without the presence of a suitable enzyme. Another reason is that enzymatic reactions can be controlled by varying the activity of the corresponding enzyme. Organisms may do that by triggering or inhibiting the transcription of the corresponding gene or by using regulatory molecules. In that way an organism is able to adapt its metabolism to a changing environment. Furthermore, such an adaption can also occur on much larger time scales. In that case, the change in the metabolic network is not initiated by an altered enzyme activity but rather by evolutionary development of the enzymes itself and the reactions catalyzed by those.

The investigation of metabolic networks therefore has been of great interest for several decades [Heinrich and Rapoport, 1974, Rapoport et al., 1976, Heinrich and Schuster, 1996, 1998, Jamshidi et al., 2001]. Parts of the metabolism, for example the glycolytic pathway, were modeled and analyzed. In recent years, a vast number of enzymatic reactions in a variety of organisms has been experimentally or computationally determined and made available

through several electronic databases. These databases contain information about the stoichiometries and to some extent also about the kinetics. With such databases, comprising thousands of reactions and metabolites, also the metabolism as a whole becomes perceptible and thus a necessity arises for concepts to analyze large scale metabolic networks.

The most accurate description of a metabolic network is certainly a dynamical model, incorporating kinetic parameters as well as enzyme and metabolite concentrations, resulting in a time-dependent description of metabolic fluxes. However, for most reactions known to occur in metabolic networks the kinetic parameters and their enzyme regulation are not known. Therefore, structural methods have been developed which can derive information from networks where only the stoichiometry is known.

These methods include algebraic methods providing information about steady state fluxes and graph theoretical approaches analyzing the topology of the networks. In the following sections selected methods will be presented in detail.

In this work a different structural method for the analysis of large metabolic networks is presented which predicts functional modules based on the network's topology. The method is based on the fact that reactions can only operate, if all of their substrates are available. Starting with some predefined resources, this condition is checked iteratively generating an expanding set of utilized reactions. This method is therefore referred to as network expansion. The set of chemical compounds which eventually can be synthesized by the expanded network has been termed "scope".

1.1 Fluxes in metabolic networks

One structural method for analyzing metabolic networks is the concept of elementary flux modes [Schuster and Hilgetag, 1994, Schuster et al., 2000]. Here, for given input and output metabolites, possible routes through the network are calculated. It is assumed that all metabolites which are not part of the set of input or output metabolites are balanced in the way that they are produced by preceding reactions at the same rate as they are consumed by succeeding reactions. This leads to a steady state where the concentrations of such internal metabolites are constant.

Generally the change of concentrations in a metabolic network is described as follows:

$$\frac{ds}{dt} = Nv, \tag{1.1}$$

where the stoichiometric matrix N is of type $c \times r$ with r being the number

of reactions and c being the number of compounds in the network. The elements of the stoichiometric matrix n_{ij} indicate the number of molecules of compound i that are consumed (negative value) or produced (positive value) by reaction j. $s = (s_0, \ldots, s_{c-1})$ defines the concentrations of the metabolites while $v = (v_0, \ldots, v_{r-1})^T$ describes the rates of the reactions which are in general dependent on s.

For the above described steady state the time derivatives of the balanced internal compounds are zero. Thus, possible steady state flux distributions converting the input metabolites into output metabolites can be calculated by

$$0 = N_I v, \tag{1.2}$$

where N_I is composed of all rows of N corresponding to the internal metabolites. v can always be equal to $\vec{0}$. If this is the only solution, no steady state flux is possible with the chosen configuration of internal and external compounds. There can also be one or more non-zero solutions v^i. In that case also $\sum a_i v^i$ is a solution of equation 1.2, where the a_i are real numbers. Hence, if the solution is non-zero, the number of possible solution vectors is infinite. To adequately describe metabolic networks it is therefore necessary, to select a finite number of characteristic fluxes. In case of reversible reactions only, a base set of the nullspace of the matrix N can be used.

In the case of irreversible reactions the situation becomes more complicated as the solution is confined to a cone in the nullspace. Schuster et al. [2000] used the concept of elementary flux modes to adequately describe fluxes in the cone. An elementary flux mode is a weighted set of reactions which can operate at steady state. The weights describe the relative flux through each of the reactions. Such an flux mode is minimal in the sense that the elimination of any of the reactions will result in an elimination of the complete flux mode. The set of all possible elementary flux modes in a network with given input and output metabolites is finite and unique. Any flux in the cone can be represented as a superposition of the elementary flux modes. The set of elementary flux modes is not necessarily linear independent.

This method is very effective for analyzing single pathways like for example glycolysis. For those the input and output metabolites are in most cases known and the number of elementary flux modes is relatively small. The method can be used to identify alternative routes. It is therefore possible to analyze, whether the network can retain its function even if some enzymatic reactions are non-functional due to a gene defect. This type of analysis can also be performed with the similar concept of extreme pathways [Price et al., 2003] which provides a set of flux vectors representing the edges of the above

mentioned cone.

In large scale metabolic networks, the number of elementary modes may become extraordinarily high, making it difficult to apply this method to this type of networks. This effect may be attenuated by considering certain highly utilized compounds as external. This may however lead to a break down of the network into smaller sub networks.

Also, for such networks it may be difficult to exactly define which metabolites are imported or exported and which are mere intermediates. A systematic approach to predict possible external compounds for arbitrary metabolic networks based on their topology is discussed in Handorf and Ebenhöh [2004].

A further method is the flux balance analysis [Bonarius et al., 1997, Edwards and Palsson, 2000, Edwards et al., 2001]. In principle, it also provides solutions of equation 1.2, however, instead of calculating a set of characteristic flux modes, it provides a single flux distribution which is optimal with respect to a predefined criterion. This criterion is dependent on the reaction rates v_i and can be formalized as follows:

$$\text{minimize } Z = \sum c_i v_i, \tag{1.3}$$

where the c_i are real numbers. For example, a particular output flux may be maximized while the input fluxes are kept small. Also, further restrictions may be applied, like keeping the the fluxes in physiological limits, $\alpha_i \leq v_i \leq \beta_i$, where α_i may be negative for reversible reactions. The solution will always be an optimal superposition of solutions of equation 1.2.

1.2 Graph representations of metabolic networks

Metabolic networks can also be represented by graphs. A graph is a mathematical object comprising of a set of nodes and a set of edges, where each edge connects a particular pair of nodes. One distinguishes between directed graphs, where edges have distinct predecessor and successor nodes, and undirected graphs, where directionality does not play a role.

The easiest way of defining such a graph is to represent all metabolites by nodes and connect all pairs of metabolites by undirected edges which take part in a common reaction [Wagner and Fell, 2001]. It is clear that such a representation looses information about which reactants actually take part in a particular reaction, if more than two metabolites are involved. A possibility to circumvent this problem is to define a so called bi-partite graph where two kind of nodes exist: metabolites and reactions. Here, metabolites

are connected to the reactions they participate in. As for all bi-partite graphs there exists no edges between nodes of the same class, i.e. there exist no edges between any two metabolites or any two reactions. The edges in the graph have to be labeled in order to define whether the metabolites are substrates or products of the corresponding reactions. Note that this classification into substrates and products shall not impose any implications on the reversibility of the reactions. The labeling rather determines the sides of the chemical reaction equation.

Recent investigations using the non-bi-partite graph representations have suggested that metabolic networks are small worlds and scale free [Jeong et al., 2000, Wagner and Fell, 2001]. Small worlds are graphs where any two nodes are connected by a path of a relatively small number of edges and which are highly clustered in the sense that different neighbors of a node have a high probability of being connected themselves [Watts and Strogatz, 1998, Strogatz, 2001]. In Wagner and Fell [2001] it was found that for a metabolic network containing 282 metabolites and 315 reactions, paths originating from the central metabolite Glutamate to all other metabolites were in average only 2.46 edges long. This demonstrates the characteristic property of small world networks.

In scale free networks, the number of nodes p_k being connected by a certain number of edges k follows a powerlaw $p_k \propto k^{-\gamma}$. The number of edges connected to a node is referred to as the degree of the node. The term "scale free" hereby accounts for the fact that the shape of the powerlaw distribution, in particular the scaling exponent γ, does not change if the abscissa is scaled by a constant factor. This is the case for the powerlaw distribution as

$$f(ck) \propto (ck)^{-\gamma} \propto c^{-\gamma} k^{-\gamma} \propto f(k) \tag{1.4}$$

Many studies have confirmed smallworldness and scalefreeness for various metabolic networks [Jeong et al., 2000, Wagner and Fell, 2001]. However, it is generally not clear, which implications may follow from it. While smallworldness is in general attributed to short paths between all nodes combined with a high robustness against removal of edges Strogatz [2001], this cannot easily be transferred to biochemical networks as the biological meaning of the edges is difficult to interpret for non-bi-partite graphs. In particular, if two metabolites in a graph are only a few edges apart, this does not mean that a synthesis of the one compound from the other needs only a few steps [Arita, 2004]. In fact, this synthesis might not even be possible.

Still, these properties can be brought in relation to the evolution of metabolic networks. Barabasi and Albert [1999] have shown, that scale free networks may emerge by preferential attachment. If new nodes are incorpo-

rated into the network, the probability of the new node being attached to an existing node increases with the degree of that node.

In a different paper [Pfeiffer et al., 2005], a model of metabolic pathway evolution was presented, where the specificity of enzymes is varied. One of the results was that it is in fact advantageous for different enzymes to share the same cofactors for specific functions. These findings apparently justify the assumption of preferential attachment. The scalefreeness of today's metabolic networks may be the outcome of such evolutionary processes.

1.3 Petri nets

Metabolic networks may also be represented as Petri nets [Reddy et al., 1996, Genrich et al., 2001, Oancea and Schuster, 2003]. A Petri net consists of places and transitions which are connected by edges. This representation is similar to the bi-partite graph representation mentioned earlier. Places represent metabolite nodes and transitions code for reactions. The edges point from places to transitions if the corresponding metabolites are substrates of the corresponding reactions. Analogously, edges point from transitions to their products. Petri nets, however, contain more information about the metabolic process than the mere topological representation. Places can contain a number of tokens which can be interpreted as the absolute number molecules or the concentration of the metabolites. Transitions can fire, which means that they transfer a certain number of tokens from their predecessor places to their successor places if a sufficient number of tokens is available. In real metabolism this represents the actual work of a reaction, converting the substrates into products. In the Petri net, the actual number of tokens taken away or put into the places is defined by the stoichiometry of the reaction. The theory of Petri nets defines invariants which can be interpreted in the context of metabolic networks [Oancea and Schuster, 2003]. There exist sets of weighted transitions, which, if executed as many times as defined by their weight, regenerate the initial token distribution, for all possible initial distributions. These reaction sets are called T-invariants and correspond to the solution vectors of equation 1.2. In fact, the calculation is done using the same methodology as before. Thus they represent steady state fluxes of the metabolic network. There also exist so-called P-invariants. These are sets of weighted metabolites for which the sum of tokens in the corresponding places multiplied by the weight remains constant for all possible combination of firing transitions. These invariants represent conserved quantities in the network. As an example, the sum $1 \cdot [ADP] + 1 \cdot [ATP]$ is constant in a network which does not include AMP nor the synthesis of

the two metabolites. It is clear, that every reaction will either not influence the concentration of the two or will convert the two into one another. Thus, without loss of generality, the number of tokens for ATP will be reduced by the same number as the number of tokens for ADP is increased. Hence, the sum remains the same.

It should be noted that conservation relations are not a specific feature of Petri nets. They can be obtained by calculating the left side kernel of the stoichiometric matrix as described in Schuster and Höfer [1991], Schuster and Hilgetag [1995]:

$$0 = cN_I \quad \text{or} \quad 0 = (N_I)^T c^T. \tag{1.5}$$

The underlying mathematics is the same as for the flux calculations (cf. equation 1.2), hence the same techniques, like elementary modes or extrem rays [Imielinski et al., 2006], may be applied.

Due to the iterative nature of Petri nets, the distribution of tokens generally will vary from step to step. When identifying the step number with time the Petri net shows time dependent behavior. It is, however, problematic to correlate this behavior to the dynamical processes in metabolism. In principle the results of a dynamical Petri net simulation are comparable to a rough numerical solution of a linear mass action kinetics, where the change of the product concentration is proportional to the product of the substrate concentrations. In the simplest form of a Petri net, as presented here, the number of tokens on the product side is increased whenever the required substrate tokens are available, i.e. if the product of their concentrations is non zero. The results can be improved by incorporating the actual substrate concentrations and the enzymatic rate constants, but these improvements will eventually just approach the solution using differential equations.

Petri nets show characteristic behavior, like deadlocks or traps whose biological meanings are discussed in Oancea and Schuster [2003], Koch et al. [2005]. However, one has to carefully separate such biological features from artifacts that simply originate from the discrete nature of Petri nets.

1.4 Biochemical databases

As mentioned above, the analysis of metabolic networks depends on the availability of biochemical information. In recent years such data became easily accessable via internet databases. The KEGG database [Kanehisa, 1997, Kanehisa et al., 2006] provides information about over 10000 chemical compounds and 6000 reactions in more than 400 organisms. Additionally, information on enzymes, genes and the corresponding annotations is available. The data is collected from various sources such as literature or other

databases. Some data is also computationally generated, like the mapping of pathway information to newly sequenced organisms which is done by comparing known enzymes sequences to the organisms genome. The Brenda database [Schomburg et al., 2000, 2004] consists of over 83000 enzymes in 9800 organisms categorized in about 4200 EC classes and acting on more than 30000 metabolites. This database also contains additional information, for example on the K_m values and inhibitors of the enzymes. Its data is extracted from literature. There exists a huge number of other resources, like BioCyc [Karp et al., 2005] or ENZYME/ExPASy [Bairoch, 2000] for metabolic networks. A list of molecular biology databases can be found in the supplement to Galperin [2006]. For this work, a non organism specific metabolic network comprising of 4811 reactions and 4104 compounds is extracted from the KEGG database. The details about the curation and modification of the data are given in appendix A.2.

1.5 Concept of Scopes

In this work a different method for the analysis of metabolic networks is used [Ebenhöh et al., 2004, Handorf et al., 2005]. The method is based on the fact that chemical reactions can only occur if all of their substrates are present. Starting with given metabolites, the seed compounds, the algorithm iteratively selects new reactions from a predefined set of possible reactions if all of their substrates are either part of the set of seed compounds or products of reactions which were already selected in an earlier iteration step. This expansion process ends when no further reactions fulfilling this condition can be found. All metabolites which can be produced by the resulting set of reactions form the scope of the seed compounds. Scopes therefore describe the synthesizing capacity of the corresponding seed compounds in a specified metabolic network.

The algorithm can be formally described as follows:

1. Selection of one or more biochemical compounds acting as a seed of the expanding network. The seed represents the first generation of the expanded network, containing an empty set of reactions.
2. Identification of those reactions from the set of possible reactions which use as substrates only compounds which are already present in the current network.
3. Incorporation of the identified reactions and their products into the network. This results in the next generation of the expanding network.
4. Repetition of steps 2 and 3 until no further reactions can be identified for

incorporation.

The above algorithm also works for reversible reactions. In such a case, a reaction can be incorporated if all substrates or all products of that reaction are present in the last network generation. Further explanations on the algorithm can be found in the appendix A.1.

After completing the process, the expanded network will contain all compounds which can be synthesized from the seed using the reactions defined in the database. This set of compounds we denote as the scope of the seed compounds. Since not all compounds can be synthesized from arbitrary seed compounds, the expansion process will in general not lead to a network containing all possible reactions.

The concept of scopes follows metabolic pathways in an intuitive way, proceeding from the substrates of a reaction to its products and further to the products of the succeeding reactions. This information can easily be obtained by looking at visual representations of biochemical pathways, like the Boehringer map.

While the benefits of the algorithm are therefore marginal for smaller networks, it is very effective for the analysis of large scale metabolic network where a visual representation is hard to obtain. Due to its low complexity, the computing times are generally very small, allowing for the systematic analysis of different seed combinations or network modifications.

The general ideas of this concept can also be found in the description of auto-catalytic sets [Kauffman, 1986, Fontana and Buss, 1994] or the chemical organization theory [Fontana and Buss, 1994, Dittrich and di Fenizio, 2007]. Their computational application becomes especially useful with the emergence of large biochemical databases.

Based on the concept discussed in this work, several papers have been recently published, including a discussion on hierarchical structuring of metabolic networks [Handorf et al., 2006], a comparison of metabolic capabilities of organism specific networks [Ebenhöh et al., 2005], a model of metabolic evolution [Ebenhöh et al., 2006], the analysis of changes of metabolic capacities in response to environmental perturbations [Ebenhöh and Liebermeister, 2006] and the prediction of possible nutrient combinations of various organisms [Handorf et al., 2007]. Further, scopes have been utilized to determine the metabolic synergy of cooperating metabolic networks [Christian et al., 2007], to predict the viability of mutant strains [Wunderlich and Mimy, 2006] and to study the effect of oxygen in metabolic networks [Raymond and Segré, 2006]. The algorithms are available in an online implementation as discussed in Handorf and Ebenhöh [2007].

1.6 Biological setting

The method of network expansion, in the way presented above, considers living cells simply as "bags of enzymes". This term describes a situation, where all necessary enzymes are present and hence, all reactions can occur as soon as their substrates are available. Furthermore, the bag is sufficiently stirred avoiding spatial differences of the chemical players. Hence, a compound available for one reaction is also available to all other reactions.

Clearly, the biological reality looks somewhat different. Cells are generally compartmented, resulting in a situation where a product synthesized in one compartment is not necessarily available as substrate to a reaction in a different compartment. The membranes between these compartments as well as the cell wall are able to let pass certain compounds while others are fixed to the compartment they have been produced in.

The concept of scopes can easily be adapted to reflect such situations. Compounds and reactions can be defined for each compartment separately. Reactions in a particular compartment transform only compounds of the same compartment. For neighboring compartments, exchange reactions can be defined transferring certain compounds across the membrane.

Still, from an evolutionary perspective, the bag of enzymes may be a good model. Certainly, any compartmentalization would also be subject to evolutionary changes. Hence, it may be useful to study the capabilities of metabolic networks without a fixed compartmentalization. A major part of this work describes a hierarchical structuring of the metabolites. In particular for this analysis it is useful to neglect compartmentalization in order to uncover the principle capabilities of the metabolism.

Further, as mentioned earlier, cells regulate their enzymes in order to adapt their metabolism to different environmental situations. Therefore, certain reactions which are in principle available in an organism may be disabled in certain situations. Also, enzymes may be expressed only in some of the compartments of the cell.

Consequently, if data on enzyme activity for various states of the cell is available, for example through microarray experiments, state specific networks can be generated and the synthesizing capacities for the different states can be analyzed [Ebenhöh and Liebermeister, 2006]. However, such data is difficult to obtain. For the work presented here, all reactions were considered to be active.

Moreover, it is possible to obtain organism specific networks. For that, reactions are considered active if they are catalyzed by an enzyme for which a corresponding gene can be identified in the organism's genome.

The question whether a reaction can occur and how fast it can transform

its substrates into its products depends on various parameters like the concentration of the metabolites and the enzyme as well as its kinetic properties. This detailed information can only be used in methods numerically solving differential equations describing the metabolic network. For this, the kinetic parameters of all participating enzymes have to be known. A more detailed comparison between the concept of scopes and dynamical modelling is given in section 1.7.

As reliable kinetic information on large scale metabolic networks is in general not yet possible to obtain, structural methods are the only way to analyze such networks. Even though the obtained results are not as accurate as the outcome of the kinetic modelling, structural methods deliver valuable insights into the metabolic capabilities of the cellular organisms.

Further, the kinetic properties of a reaction together with the metabolite concentrations determine the direction in which a reaction will proceed. In principle, for each reaction, metabolite concentrations can be chosen to force the reaction to proceed in one or the other direction. However, under normal physiological conditions, metabolite concentrations are generally bound to a certain interval. Therefore, if the kinetic parameters are suitable, certain reactions will always proceed in only one direction.

In that way it is possible to integrate precalculated kinetic information into structural methods by allowing certain reactions to be used only in a predefined direction. However, apart from the fact that the necessary information may not be present for all reactions, this information may be also misleading. For analyses in evolutionary context or robustness studies, the assumption that the metabolites have still the same physiological concentrations may be inaccurate. Consequently, the information on the reversibility of the reactions may become misleading.

In this work, most calculations have been performed assuming all reactions as reversible. Section 4.5 describes the changes to the results if information on reversibility is included in the model.

Generally, living cells take up resources and synthesize consecutively various intermediates and eventually the desired final products. Such products may be exported to the extracellular medium and may include metabolites needed by other tissues in multicellular organisms, toxins or by-products. Further, a major role of the cell's metabolism is the production of compounds needed for cell growth and division. Such products are often referred to as biomass. Even though the distinct modelling of cell growth and division is far beyond the scope of this work, such processes can be considered by interpreting growth as a dilution of all metabolites in the cell. In effect, it has to be assumed that metabolism has to continuously refresh all its metabolites, including all intermediates.

As described, a scope consist of all compounds that can be produced from the seed. This implies that there exist metabolic fluxes converting the seed into all the other compounds in the scope. It should however be noted that this does not imply a steady state of this metabolic flux. It only assures that compounds in the scope will be produced in an initial transient phase when an empty network is provided with the seed. Compounds outside the scope will not be produced, neither transiently nor in steady state. The following two sections will deal with this fact in more detail.

1.7 Comparison to dynamical models

The most exact results in analyzing metabolic networks can be obtained when considering the kinetics of the participating reactions. One way to incorporate such knowledge is the utilization of differential equations. A reaction system can be described by the differential equation 1.1. The reaction rates are in general dependent on the concentrations of the participating compounds. To reflect this, the equation can be rewritten as:

$$\frac{ds}{dt} = Nv(s), \tag{1.6}$$

with N being the stoichiometric matrix, s the vector of metabolite concentrations and v the vector of reaction rates.

$v_i(s)$ describes the kinetics of reaction i converting the substrates C into the products P

$$C \rightleftharpoons P. \tag{1.7}$$

Generally, the kinetic of a reaction follows the law of mass action, stating that the rate of a reaction is proportional to the product of its substrate concentrations. As chemical reactions are reversible, the effective reaction rate is the difference of the forward rate, describing the transformation of the substrates into products, and the backward rate, describing the reverse reaction. In the following, this rate is called v without the index i as only one reaction is considered. Hence, the reaction rate v can be expressed as follows:

$$v = k_+ c_1 \cdot \ldots \cdot c_l - k_- p_1 \cdot \ldots \cdot p_m, \tag{1.8}$$

where the c represent the concentrations of the l substrates and the p the concentrations of the m products of the reaction. The two constants k_+ and k_- depend on steric properties of the participating compounds and the reaction site as well as the energy of the compounds and the transition state. In general these constants can be determined experimentally.

For the majority of the reactions occurring in living cells, the kinetic constants k_+ and k_- are so small that they could be completely neglected. However, in metabolism, such reactions are generally catalyzed by enzymes which can increase the reaction rates by many orders of magnitude. This effect is achieved by first modifying the steric arrangement of the reactants in order to perform the transformation and second by lowering of the energy of the transition state.

Clearly, the binding of the substrates to the enzyme can again be described by mass action. Equation 1.8 can be modified in the following way:

$$v = \tilde{k}_+ \hat{e} \prod_k c_k - \tilde{k}_- \hat{e} \prod_k p_k, \tag{1.9}$$

Here, the $\tilde{k}_+$ and $\tilde{k}_-$ reflect the different, much faster, kinetics of the reaction with the enzyme present. $\hat{e}$ is the concentration of the enzyme. Hence, the reaction rate is linearly increased with increasing enzyme concentration. However, in living cells, the number of enzyme molecules is limited. Therefore, a significant part of the enzymes may be in use by the reaction itself. Thus, the free enzyme concentration e is actually dependent on the other variables in the system. This problem can only be solved by considering the free concentration e of the enzyme E as a dynamical variable of the system and regarding substrate binding and product release as separate reactions steps:

$$E + C \underset{k_{1-}}{\overset{k_{1+}}{\rightleftharpoons}} EC \underset{k_{2-}}{\overset{k_{2+}}{\rightleftharpoons}} E + P \tag{1.10}$$

and

$$v_1 = k_{1+} e \prod_k c_k - k_{1-} z \tag{1.11}$$

$$v_2 = k_{2+} z - k_{2-} e \prod_k p_k \tag{1.12}$$

Here, z represents the concentration of the enzyme-substrate-complex. If the two reactions proceed in a faster time scale than the changes in the metabolite concentrations of the substrates C and the products P, a quasi steady state approximation for z can be used:

$$\frac{dz}{dt} = v_1 - v_2 = 0. \tag{1.13}$$

The reaction rate of the complete reaction v can be written as (see ap-

pendix A.4 for a derivation):

$$
\begin{aligned}
v &= \frac{\frac{V^+_{max}}{K^+}\prod_k c_k - \frac{V^-_{max}}{K^-}\prod_k p_k}{1 + \frac{\prod_k c_k}{K^+} + \frac{\prod_k p_k}{K^-}}, \\
V^+_{max} &= \hat{e}k_{2+}, \quad V^-_{max} = \hat{e}k_{1-} \\
K^+ &= \frac{k_{1-} + k_{2+}}{k_{1+}}, \quad K^- = \frac{k_{1-} + k_{2+}}{k_{2-}}
\end{aligned}
\tag{1.14}
$$

Hence, as the total enzyme concentration $\hat{e}$, i.e. the free form e plus the bound form z, is independent of the reaction rate, the reaction can be modelled without considering the enzyme concentration as a dynamical variable. However, instead of the linear mass action kinetics (1.8) the kinetics given in 1.14 has to be used.

In the case of a rapid drain of the products ($p_k \to 0$), the second reaction can be considered as effectively irreversible. Then equation 1.14 is transformed into the Michaelis-Menten kinetic:

$$
v = \frac{V^+_{max}\prod_k c_k}{K^+ + \prod_k c_k} \tag{1.15}
$$

The parameters V_{max} and K can be determined experimentally [Stryer, 2003].

In many cases, if the concentrations of the metabolites and enzymes are large enough to justify modelling with continuous variables and if further inhibitory or activating processes can be neglected, solutions of the differential equation system 1.6 describe metabolic systems very accurately. In its simplest form, also here a "bag of enzymes", as described in section 1.6 is assumed. Analogously, also differential equation systems can easily be extended to compartmented models by the inclusion of transport reactions. Further, an extension to partial differential equations is possible, allowing for gradients in the metabolite concentrations within the cell.

The expansion algorithm used for the calculation of scopes approximates the law of mass action. As described above, a reaction can only be incorporated in the expanding network, if all its substrates are present, i.e. having a non zero concentration. Consequently, the products of this reaction will also be added to the expanding network. Analogously, in the system of differential equations, a reaction rate is non-zero, if the product of the substrate concentrations is non-zero. After a finite time, this will result in non-zero concentrations of the products. Clearly, this also holds for the case of an enzymatic reaction if the enzyme concentration is non-zero.

If the initial conditions of the differential equation system are chosen in a way that the seed compounds have a finite concentration and all other compounds have a concentration of zero, the set of compounds that have a non-zero concentration after a sufficient period of time will coincide with the set of compounds defined by the scope of the seed compounds.

This equality justifies the term synthesizing capacity for the scope. Even though the concept of scopes cannot provide distinct values for concentrations nor time courses, it is able to provide functional modes of metabolic networks for arbitrarily chosen external resources. These modes give frames for the actual time dependent behavior of the network as described by differential equations.

Due to the small computing times of the scopes, the concept allows for the systematic analysis of resource combinations, variations of the network structure or cross-species comparisons, which more sophisticated methods are not able to deliver.

1.8 Comparison to other structural methods

The concept of scopes extends the graph theoretical analysis by stoichiometric constraints. In fact, with this method it is only possible to traverse from a substrate of a reaction to its product if all other substrates are also available. Certainly, paths through a metabolic network calculated with this method will differ dramatically from those calculated using less restrictive graph representations.

The stoichiometric constraints put on the graph traversal are actually similar to the function of Petri nets. However, the method of network expansion does not intend to simulate time dependent behavior. In fact, the method resembles a Petri net, where once a compound has gotten a token it cannot loose it anymore. The method is therefore not subject to typical dynamical behaviors like oscillations or deadlocks.

Flux based methods, like elementary flux modes, extreme pathways or flux balance analysis, predict metabolic fluxes, generally converting external input metabolites into external output metabolites via a number of balanced intermediates, as discussed in section 1.1. Such steady state fluxes are described by equation 1.2. The input and output metabolites can be explicitly integrated in this equation:

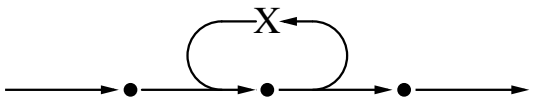
$$\begin{pmatrix} 0 \\ a_i \\ b_j \end{pmatrix} = \begin{pmatrix} N_I \\ N_{E\uparrow} \\ N_{E\downarrow} \end{pmatrix} v, \qquad a_i \leq 0, b_j \geq 0. \tag{1.16}$$

Here, N_I is the part of the stoichiometric matrix representing the balanced internal compounds, $N_{E\uparrow}$ represents possible input metabolites (comparable to the seed of the network expansion) and $N_{E\downarrow}$ possible output metabolites.

The method of network expansion assumes a network where all metabolites except the seed compounds have initially zero concentrations. If after an initial transient phase the reaction network settles in an equilibrium, steady state fluxes may exist. These fluxes use the seed compounds as input metabolites and may use all other compounds in the scope as intermediates or output metabolites. Compounds outside the scope will not be affected by any of these fluxes and therefore have constant zero concentrations.

In the case of cell growth as discussed in section 1.6 all metabolites can be assumed as output metabolites. Under that assumption there may exist steady state fluxes through all reactions in the expanded network. The reason is that each compound in the scope can be produced at steady state if the substrates of the reaction producing that compound can be produced at steady state. Possible additional products of this reaction will consequently be output metabolites. Clearly, the same is true for the steady state production of the substrates. Hence, the expansion process can be traced back to the seeds to confirm that all metabolites are produceable from the seeds at steady state.

Compounds not included in the scope can in general not be produced if no other substrates are provided. There may however exist compounds which are required by some reactions but will be produced only in a later step:



Whereas flux based methods will automatically predict products whose synthesis requires the presence of X, the network expansion will stop after the first reaction step if X is not explicitly added to the seed.

In a living cell, such situations are however not common. Such a compound X will eventually vanish through degradation or dilution. Even if the compound is produced in a larger amount to compensate for this, the cell would have only a weak influence on the concentration of X, which may result in a loss of control on the whole pathway. Therefore, if such a reaction path is essential for the cell, it can be assumed that there exists a way to directly produce X.

However, there exist two occasions where such a situation may have a biological relevance. First, the non-linear autocatalytic effect may be desired. This case, however, is more known to occur in signal transduction networks

rather than in metabolism. Second, compound X may be provided or regulated by other parts of the cell, which are not considered in the model under investigation. This is often the case for cofactors, which are assumed to be present when analyzing metabolic subsystems. In this case the cell would still have the control on the concentration of X and hence, the control on the path.

Indeed, the synthesis of many cofactors are complex processes. Such a process generally requires the presence of cofactors, in certain cases also the presence of the cofactor to be produced. This is not a problem if these cofactors are ubiquitous and their concentrations are kept on a relatively constant level by other regulatory systems of the cell. For example, the synthesis of ATP in glycolysis first consumes two molecules of ATP before producing 4 ATP molecules in the end. Under physiological conditions, the homeostatic regulation of ATP (Rapoport et al. [1976]) holds the concentration of ATP on an approximately constant level. This avoids situations where the ATP concentration is so small that its synthesis is inhibited.

The concept of scopes can capture such situations through special treatment. Therefore, certain cofactor functionalities are assumed to be present in the network while avoiding that the cofactors themselves are used as substrates for the synthesis of other metabolites. Details can be found in section 2.5.

Apart from the above mentioned difference, the scopes and flux modes also vary in other aspects. While flux modes describe potential reaction routes between predefined input and output metabolites, scopes represent only one functional module which describes the answer of the metabolism to a specific set of input compounds. While the flux based methods are the best choice for obtaining possible steady state fluxes for predefined input and output metabolites, the scopes provide a good measure for the metabolic capability of a network when certain resources are available in the environment. The scope can be interpreted as a flux mode which uses the seed as input metabolites and all other compounds as outputs.

1.9 Properties of Scopes

The set of compounds which are contained in the expanded network resulting from a single seed compound A, we denote by $\Sigma(A)$ and call it the scope of A. By the scope size $\sigma(A)$ we denote the number of compounds contained in the scope $\Sigma(A)$. Corresponding to the set of compounds, the final network also contains an associated set of reactions denoted $W(A)$.

Clearly, if a compound B is included in the scope of A, then the scope of

B is a subset of the scope of A, formally:

$$B \in \Sigma(A) \quad \text{is equivalent to} \quad \Sigma(B) \subseteq \Sigma(A). \tag{1.17}$$

Further, if two compounds A and B are interconvertible in the sense that A can be produced from B and B can be produced from A (without using other compounds as substrates), then A is included in the scope of B and B is included in the scope of A. This implies that the scopes of A and B are identical, formally described by:

$$B \in \Sigma(A) \text{ and } A \in \Sigma(B) \quad \text{is equivalent to} \quad \Sigma(A) = \Sigma(B). \tag{1.18}$$

There exist nesting in the sense that if a compound B is in the scope of A and a compound C is in the scope of B then C is also in the scope of A:

$$B \in \Sigma(A) \wedge C \in \Sigma(B) \Longrightarrow C \in \Sigma(A). \tag{1.19}$$

Scopes may also be defined for a seed consisting of multiple initial compounds $A_1, \ldots, A_k$. This results in the so called multi scope $\Sigma(A_1, \ldots, A_k)$. If it is necessary to distinguish between scopes of a single seed compounds and scopes of multiple seed compounds, the terms 'single scope' and 'multi scope' will be used. Equations 1.17 to 1.19 analogously hold for multi scopes.

It is evident that a multi scope cannot be smaller than the union of the single scopes $\Sigma(A_1), \ldots, \Sigma(A_k)$ of the individual compounds.

$$\Sigma(A_1, \ldots, A_k) \supseteq \Sigma(A_1) \cup \ldots \cup \Sigma(A_k) \tag{1.20}$$

The symbol Σ can be seen as an operator mapping a set of compounds to a new set of compounds, the scope. Σ is a projection operator which is idempotent:

$$\Sigma(\Sigma(S)) = \Sigma(S) \quad \text{or} \quad \Sigma^2 = \Sigma \tag{1.21}$$

Hence, a set of compounds S is a scope if the following condition holds:

$$\Sigma(S) = S \tag{1.22}$$

Equations 1.21 and 1.18 also indicate that a seed is always interconvertible with its scope.

Further, the cut set of two scopes is a scope:

$$\Sigma(\Sigma(S_1) \cap \Sigma(S_2)) = \Sigma(S_1) \cap \Sigma(S_2) \tag{1.23}$$

Proof:

$$C = \Sigma(S_1) \cap \Sigma(S_2) \tag{1.24}$$

Let Z be the scope of C

$$Z = \Sigma(C) \tag{1.25}$$

then, with equation 1.17,

$$Z \subseteq \Sigma(S_1) \wedge Z \subseteq \Sigma(S_2). \tag{1.26}$$

Consequently, Z must be part of the cut set of S_1 and S_2:

$$Z \subseteq C. \tag{1.27}$$

which means that Z equals C as a scope cannot be smaller than its seed.

Chapter 2

Scopes of Compounds

2.1 Scopes of single compounds

In the following, the concept of scopes has been applied to the metabolic network retrieved from the KEGG database (cf. Appendix A.2). In particular the single scopes of 4104 compounds have been calculated. Due to it's ubiquity, water is assumed to be present for all calculations in this work, unless otherwise stated. Methodically this means, that water is added to all seeds. Despite of the fact that there are 2 compounds in each seed, these scopes will still be termed as single scopes. Using the available reactions, water itself can be transformed into 4 other metabolites, namely O_2, H_2O_2, H^+ and O_2^-. All scopes therefore contain at least these 5 metabolites.

Figure 2.1 shows the distribution of the scope sizes. The sizes range from 5 to 2183 compounds. The distribution is found to be very non-uniform. While most of the scopes are rather small, there also exist a few large scopes. Furthermore, for sizes larger than 32 the distribution contains gaps, which may become very wide between larger scopes. There exists only a small number of very large scopes, in particular with the sizes 1554, 1556, 1558, 1560, 1596 and 2183. The next smaller scope has only a size of 560.

As expected, some of the compounds result in the same scope (cf. equation 1.18) which can be seen as large peaks in the distribution. In fact, there exist only 2923 distinct scopes. To demonstrate the effect on the distribution, figure 2.2 shows the scope sizes of the distinct scopes only. It can be seen that for small scopes typically several distinct scopes with the same sizes exist whereas for larger scopes the large number at a certain size in figure 2.1 is mainly determined by interconvertible seed compounds.

Table 2.1 lists the largest single scopes sorted by size. The largest scope of size 2183 results from four different single compound seeds which are

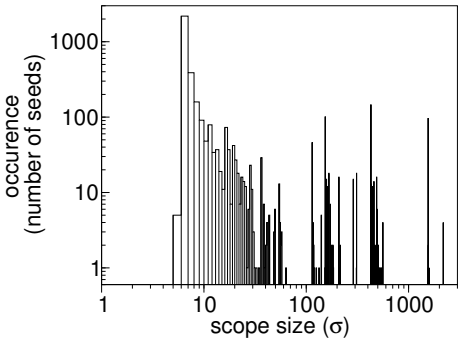


Figure 2.1: Size distribution of the scopes of 4104 single compounds.

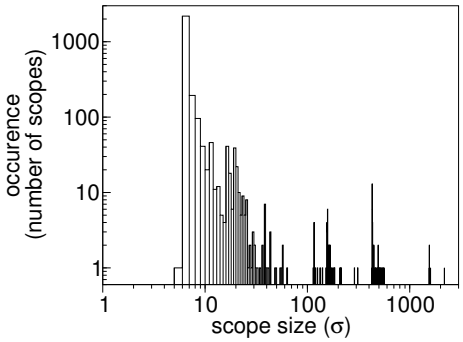


Figure 2.2: Size distribution of the 2923 distinct scopes.

adenosine 5'-phosphosulfate (APS), 3'-phosphoadenosine 5'-phosphosulfate (PAPS), dephospho-CoA, and UDP-6-sulfoquinovose. APS and PAPS play an important role in the sulfur metabolism in many microorganisms. Dephospho-CoA is a direct precursor in the CoA biosynthesis pathway. UDP-6-sulfoquinovose plays a role in the glycerolipid metabolism.

Of particular interest is also the scope of size 1554 which can be reached from 97 different single seed compounds. Among them are central cofactors such as ATP, UTP, CTP and GTP as well as the corresponding mono- and diphosphates and the nicotinamide dinucleotides NADH and NADPH.

Many scopes are subsets of larger scopes (see equation 1.17). For example, the scope of ATP is a subset of the scope of APS. From this it follows that ATP can be synthesized from APS. The opposite process is not possible which simply follows from the fact that APS is composed of adenine, ribose, sulfate and phosphate, whereas the adenosine phosphates AMP, ADP, and ATP contain the same building blocks except sulfate.

2.2 Interconvertibilities

As shown, many biochemical compounds are interconvertible as described by equation 1.18. Obviously, it is necessary for two interconvertible compounds that they are composed of the same chemical elements. However, not all compounds fulfilling this condition are interconvertible. This is the case if the reactions present in the network do not have the capability to perform the interconversion. As an example, we consider the two compounds coenzyme A and dephospho-CoA. Both substances consist of the same chemical elements. The only difference is that coenzyme A contains three phosphate groups

KEGG ID	compound name	scope size
C00053	3'-Phosphoadenylyl sulfate	2183
C00224	Adenylylsulfate	2183
C00882	Dephospho-CoA	2183
C11521	UDP-6-sulfoquinovose	2183
C00016	FAD	1596
C04652	UDP-2,3-bis(3-hydroxytetradecanoyl)glucosamine	1560
C06435	5'-Butyrylphosphoinosine	1558
C05227	UDP-sugar	1556
C01299	Adenylyl-[L-glutamate:ammonia ligase (ADP-forming)]	1556
C00002	ATP	1554
C00003	NAD	1554
C03483	Adenosine 5'-tetraphosphate	1554
⋮	⋮	⋮

Table 2.1: List of single compound seeds and their scope sizes ordered by decreasing size (abbreviated). Corresponds to the largest scopes shown in figure 2.1

whereas dephospho-CoA contains only two. Our calculations revealed that coenzyme A is in the scope of dephospho-CoA, whereas the opposite is not true. Even though the network includes the reaction

$$\text{dephospho-CoA} + \text{ATP} \rightleftharpoons \text{CoA} + \text{ADP}, \tag{2.1}$$

it does not represent a direct interconversion between the two compounds since it requires the presence of ATP or ADP. However, coenzyme A can be produced from dephospho-CoA in a higher number of steps. This is possible since ATP is in the scope of dephospho-CoA (see above). In contrast, dephospho-CoA cannot be produced from coenzyme A since its scope does not contain ADP.

In order to get an impression of how many of the compounds containing the same elements are really interconvertible, we have analyzed all pairs of compounds containing only the elements C, H, and O. The database contains 1501 such compounds forming 1125750 different pairs. From these pairs of compounds, only 6126 pairs (0.54%) represent two compounds which can be interconverted. Analogously, we have analyzed all 186 compounds containing the elements C, H, O, N, P, and S. It turns out that 1.24% of all pairs of these compounds are interconvertible. Interestingly, for all 363 compounds containing the elements C, H, O, N, and P, over 7% of all pairs are interconvertible. This high percentage can be explained by the fact that many

of those compounds are seed compounds of the scope of ATP. Table A.1 in section A.5 summarizes the results for all existing element combinations.

2.3 Multi scopes of small building blocks

As shown above, the analysis of scopes of single compounds can yield interesting information about the analyzed metabolic network. However, it is not very realistic that a real metabolic network is actually supplied with such relatively complex compounds like ATP or APS. In fact, it is more realistic to assume that the seed contains several less complex compounds, from which the more complex metabolites are eventually synthesized.

As previously shown, APS as well as the other three compounds PAPS, dephospho-CoA, and UDP-6-sulfoquinovose possess the largest scope. These compounds are rather complex and produced by intracellular processes. For example APS is produced by the enzyme sulfate adenylyltransferase, converting ATP and sulfate into APS and pyrophosphate. It is an intriguing question, whether scopes of similar sizes can be obtained when starting the expansion process with a small number of simple compounds which can be assumed to be present in the environment. Guided by the elements contained in APS (see previous section), the following set of seed compounds is chosen: CO_2, NH_3, H_3PO_4 and H_2SO_4. Interestingly, the set of compounds which can be synthesized from these simple compounds is exactly the same as the set produced from APS. Starting the expansion process only with the building blocks CO_2, NH_3, and H_3PO_4, i.e. omitting the sulfur source, results in a multi scope which is identical to the scope of ATP, indicating that the two seeds are interconvertible. In other words, ATP can be produced from CO_2, NH_3, and H_3PO_4 and these, in turn can be produced from ATP.

We have further tested whether the scopes remain the same when replacing the carbon source CO_2 by CH_4. The resulting scopes posses size 25 and 19, for the case with and without sulfate, respectively, containing predominantly inorganic compounds. In both cases, the small scope sizes are due to the fact that all reactions utilizing methane require the presence of cofactors like NAD^+ in a very early stage. A detailed discussion of the role of the cofactors is given in section 2.5.

The scope of a set of compounds which is proposed as hypothetical inorganic precursors for the origin of life (Martin and Russell [2003]), namely H_2CO_3 (carbonic acid), CH_3SH (methanethiol), NH_3 and $P_2O_7^{4-}$ (pyrophosphate) is again identical to the scope of APS, the largest single scope of the complete network. Extending the seed by CO_2, CH_4, and CN^- (cyanide), compounds which are also discussed in Martin and Russell [2003], does not

further increase the scope size.

2.4 The expansion process

Not only the scope itself, but also the expansion process leading to it can reveal interesting features of the metabolic network. During that process the "discovered" part of the network grows by incorporation of new reactions and compounds. The analysis of the expansion is most interesting for larger scopes as larger parts of the network and thereby potentially interesting features are traversed. Here, the expansions starting with APS and the set CO_2, NH_3, H_3PO_4, and H_2SO_4 are analyzed.

Figure 2.3 shows the number of reactions and compounds over the network generation during the expansion process of APS. The numbers of compounds and reactions increase slowly in the beginning, faster in the middle phase and saturate in the last phase.

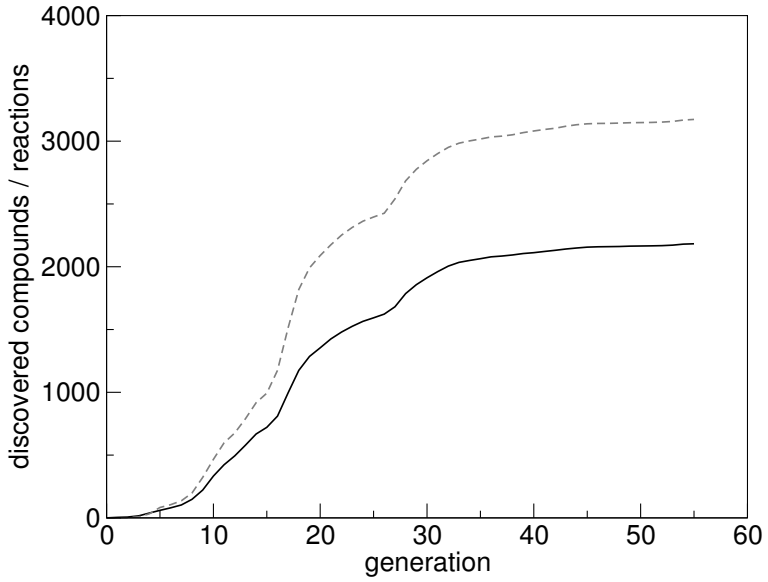


Figure 2.3: Expansion process starting with APS as seed. Shown are the number of compounds present in the discovered part of the network in a certain generation (solid line) and the number of the corresponding reactions (dashed line).

This behavior can be expected from a network without special structural features: in the beginning the discovered network is small and consequently there are only a few compounds which can serve as substrates for the reactions to be incorporated. Therefore only a few reactions can be added, even though

there are many reactions still available. In the second phase the network has grown to such a size that it contains many compounds which can be used as substrates in the further development. Also, the set of still available reactions is considerably large. This results in a strong increase of the network size in this phase. Towards the end of the process the number of reactions still available for the process becomes smaller. Therefore only few reactions can be added per generation which results in a final saturation phase. These considerations can be summarized in a theoretical model of the expansion process and are discussed in the appendix in section A.6.

A closer inspection of the expansion curve in figure 2.3 reveals that on top of the mentioned sigmoidal structure there exist phases of temporal slow-down and acceleration in the expansion. These may be the result of special topological features in the network which are traversed during the expansion process. They become even more apparent when looking at the number of compounds and reactions which are added per generation as shown in figure 2.4.

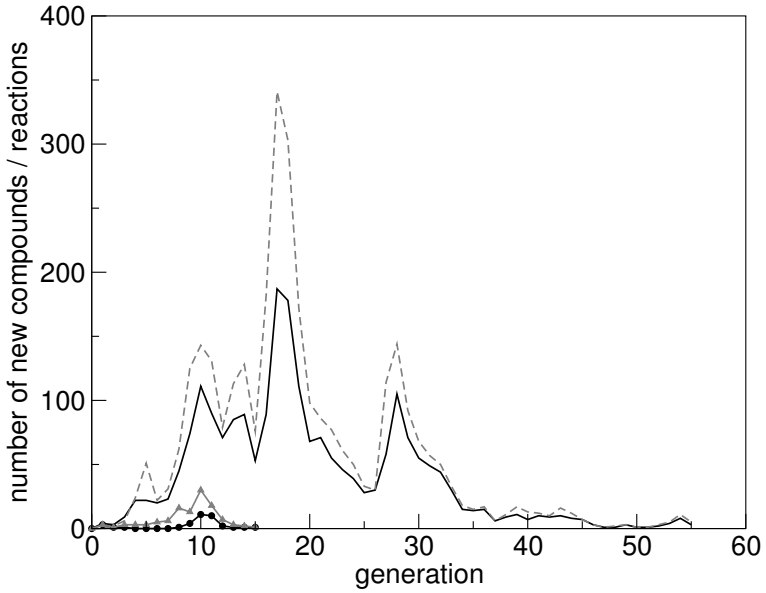


Figure 2.4: Expansion process starting with APS as seed. Shown are the compounds added in a certain generation (solid line), the corresponding reactions (dashed line), the fraction of compounds present in the glucose scope (circles) and in the glucose-phosphate scope (triangles).

The expansion curves suggest that at critical stages certain key compounds are incorporated which drastically accelerate the expansion of the network. In the following, such critical stages are analyzed in closer detail.

The phenomenon of temporary slow down of the expansion is most predominant at generations 7, 15 and 26. The first 7 generations are characterized by disassembly of APS into compounds like AMP, ribose phosphate, adenosine, sulfate and phosphate as well as related compounds like ATP, IMP, ITP, NH_3 and so on. With the inclusion of fructose-6-phosphate, CO_2 and glycerone phosphate in generation 7 and pyruvate in generation 8 the expansion is strongly accelerated. In subsequent generation a variety of sugars and other carbohydrates are synthesized. In fact, many compounds of this phase can be found in the scopes of glucose and glucose phosphate as shown in figure 2.4.

Between generations 10 and 15 the number of still available reactions which can utilize the sugars and sugar phosphates decreases, leading to a slow-down of the expansion process. The situation changes dramatically in generation 15 when NAD^+ is incorporated into the expanding network. This compound is capable of acting as a cofactor in a large number of reactions. In fact, NAD^+ participates in 147 of the 182 new reactions added in generation 16. The expansion process is further accelerated in generation 16, mainly by incorporation of the compounds NADH and $NADP^+$. The exploitation of the capabilities of the nicotinamide dinucleotides to act as cofactors leads to a deceleration of the expansion process after generation 17. Similarly, the incorporation of coenzyme A in generation 26 activates a huge amount of CoA-dependent reactions in the subsequent generation. After the majority of these reactions is incorporated the expansion slows down again. As a tendency, this deceleration continues until the end of the expansion process as most reactions included in the scope of APS are already incorporated.

During the whole process, the number of reactions added per generation is approximately proportional to the number of compounds. Generally, there are more reactions added per generation than compounds. This means that incorporation of a new reaction does not necessarily allow for the synthesis of additional compounds. This indicates the existence of alternative pathways for the production of certain compounds, as some newly added reactions only produce compounds which already have been produced by other reactions in previous steps.

As a second example, the expansion starting from CO_2, NH_3, phosphate and sulfate is considered. The scope is identical to the scope of APS, whereas the expansion processes shows some differences as indicated in Figure 2.5.

In both cases, the expansion proceeds slowly at the beginning as well as near the end of the process. In between there are phases of deceleration and subsequent acceleration. However, in the case of APS, the scope is reached after 55 generations whereas in the case of the small building blocks 61 generations are required.

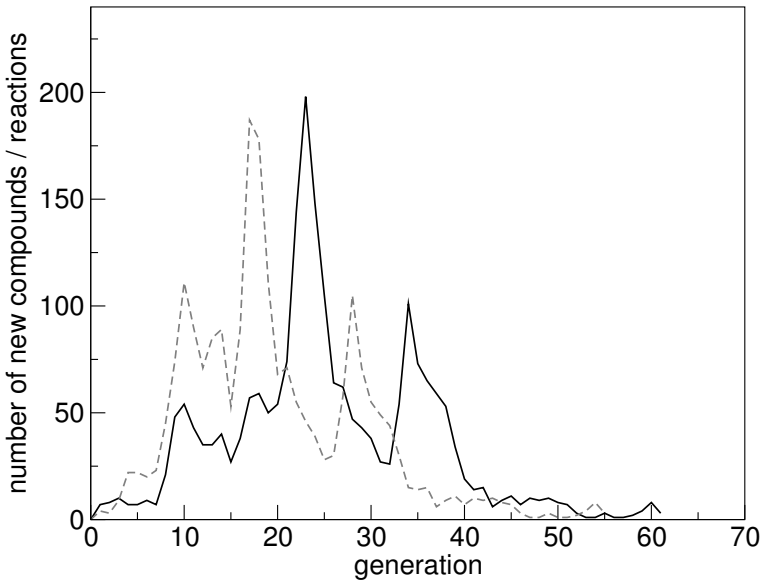


Figure 2.5: Expansion process starting with CO_2, NH_3, phosphate and sulfate as seed (black line) in comparison to the expansion starting with APS (dashed line). Shown are the numbers of compounds added per generation.

Both curves show three pronounced peaks. The first peak, appearing in both curves around generation 10, can be explained by the incorporation of the majority of carbohydrates. In both processes, the second strong increase in the number of incorporated compounds results from the incorporation of NAD^+, NADH, $NADP^+$ and NADPH, as well as ATP and ADP in the case of the four building blocks, into the network. Notably, in the expansion process starting with the four simple compounds this peak is shifted towards later generations. This can be understood by the fact that the adenine nucleotides, which are also important precursors in the synthesis of NAD^+ and NADP can simply be extracted from APS while they have to be synthesized from the simple compounds in a larger number of reactions. After the second peak the two expansions proceed in an almost identical way. From that we conclude that in this stage both expansion processes, one beginning with a single complex compound and the other starting with four simple molecules, have reached almost the same subnetwork. Consequently, in both processes the numbers of generations between the second and the third peaks, corresponding to the inclusion of CoA, are the same.

2.5 The role of cofactors

As shown in the last section, the expansion processes strongly accelerate after the appearance of the cofactors NAD^+/NADH, $NADP^+$/NADPH, ATP/ADP and coenzyme A (see Figures. 2.4 and 2.5). One reason is certainly the large number of reactions in which each of these cofactors takes part. This, however, cannot be the only reason, since the effect of ATP is only visible in the expansion starting from CO_2, NH_3, phosphate and sulfate, but not in the expansion starting from APS. As the name cofactor already suggests, these substances are additional requirements to reactions metabolizing other compounds. Therefore, the cofactor alone cannot have a large effect on the expansion if these other compounds are not present. In the case of APS, ATP is produced very early, when only a few other compounds are already present. Consequently, ATP cannot use its phosphorylating capability to cause a strong acceleration effect.

To summarize, the peaks can be explained such that before a peak there exist a variety of compounds waiting for being phosphorylated, oxidized, reduced or for a transfer of another chemical group. With the incorporation of the corresponding cofactor all these reactions become active and will be included in the subsequent generations.

Scopes can be calculated under the assumption that these cofactors are present. This assumption complies with the biological reality, where analyzed cells which are going to be fed with external substrates (seed) are not empty but contain the whole set of essential metabolites. However, it has to be ensured that the cofactors are only used in their function as cofactors and not as substrates for the synthesis of other metabolites in the scope.

The expansion process has to be modified in order to reflect these consideration. In the case of NAD^+, 562 of the 573 reactions in which NAD^+ takes part are also connected to NADH. The cofactor pair NAD^+/NADH oxidizes or reduces other metabolites by taking up or donating a proton. This functionality is only present in a reaction if NAD^+ and NADH are present on different sides of the reaction with the same stoichiometries. Otherwise, the corresponding reaction takes part in the synthesis or degradation of NAD^+ or NADH. In this way, the modified algorithm can detect reactions in which the metabolites operate as cofactor and can incorporate them even though the cofactors are not produced in the expansion process itself. More details are given in the appendix in section A.3.

It is an intriguing question in which way the expansion process is affected if the functionality of certain cofactors is present in the network. Redox reactions are often catalyzed by NAD^+, $NADP^+$, NADH and NADPH. The modified expansion process allows the incorporation of all reactions which

would otherwise require the presence of these cofactors. Such a process, starting with the seed compounds CO_2, NH_3, H_3PO_4 and H_2SO_4 is depicted in figure 2.6a (solid line). Since the original expansion (dashed line) includes the synthesis of these cofactors, the modified expansion results in the same final network but in a smaller number of generations. As expected, the strong increase in the expansion's velocity, which in the original process is invoked by the appearance of NAD^+, now takes place much earlier. There is, however, an initial lag phase of about 7 generations, which is required for the synthesis of those compounds participating in the corresponding redox reactions.

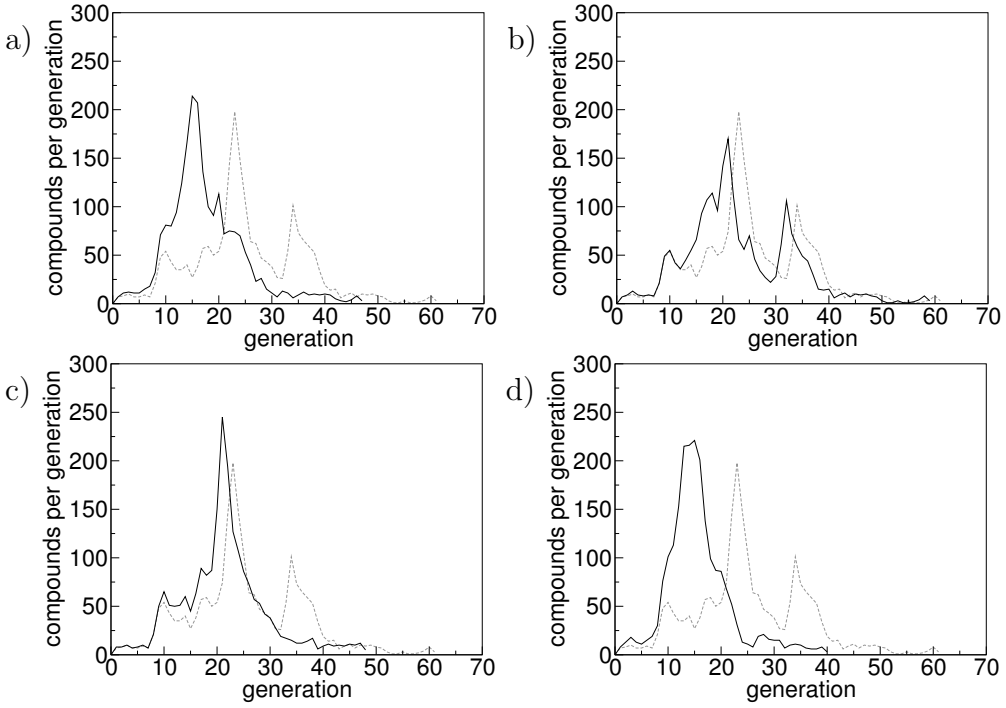


Figure 2.6: The effect of cofactors on the expansion process starting from CO_2, NH_3, H_3PO_4 and H_2SO_4: a) The cofactors NAD^+/NADH and $NADP^+$/NADPH are present; b) The cofactors ATP/ADP; c) Coenzyme A; d) All mentioned cofactors together. The dashed line always represents the unmodified process

Other cofactor functionalities also affect the expansion process. For example, phosphorylations which normally take place at the expense of ATP can now occur without that ATP is explicitly provided. This leads to the modified expansion process depicted in figure 2.6b. Clearly, the modification results in an accelerated expansion but the effect is less pronounced than in the case of the replacement of the cofactor NAD^+. Acyl groups may be

transferred even without the explicit presence of coenzyme A. The resulting process is depicted in figure 2.6c. Again the process is accelerated and, as expected, the peak resulting in the original process from the incorporation of CoA disappears.

Further, the case where the functionalities of the four cofactors are available is analyzed. Not surprisingly, this combined modification results in a process (figure 2.6d) which is faster than any process obtained by addition of a single cofactor. Whereas the original expansion process, and to a less extent also the modified processes using single cofactors, show phases of temporary slow down, this property is in principle no longer observed in the case of combined modification. Instead, the process accelerates continuously until generation 15 and subsequently decelerates.

Starting with other seed compounds which in the original process do not produce the considered cofactors, the addition of the functionalities of the cofactors will generally lead to an increased size of the scopes. The corresponding effects can be dramatic. For example, the original scope size of CO_2 is 17. Adding the functionalities of NAD^+, $NADP^+$ and CoA leads to a resulting network which contains 686 compounds. This effect can be analyzed more systematically. In figure 2.7 the scope size distribution of all single seeds is shown for expansions without a) and with b) the cofactors ATP/ADP, NAD^+/NADH, $NADP^+$/NADPH and CoA. Clearly, with cofactors, the scope sizes are generally increased. There exists now a large gap between size 50 and 686. This can be explained by the fact that, with the help of the cofactors, a large number of compounds containing the elements C, H and O can be synthesized from simple carbon containing metabolites like CO_2, methane or formate.

This effect strongly underlines the importance of cofactors in biological reaction networks. Even though many compounds contain the necessary elements to synthesize a large number of products, the present chemical reactions are not able to do the corresponding transformations (see section 2.2). With the presence of cofactors many of the formerly impossible transformations can occur, resulting in grossly increased scopes sizes for a large number of compounds, as shown in figure 2.7b.

The so far analyzed cofactors have been selected due to their biological importance which is well-founded by the large number of reactions catalyzed by these cofactors. In principle this set can be extended to other cofactors using biological knowledge.

In this work, however, a different approach is used. By looking at the topology of the reaction network possible cofactors are predicted. The basic assumption is that cofactors appear in pairs which share a common base structure and only differ in a functional group. The two cofactors of a co-

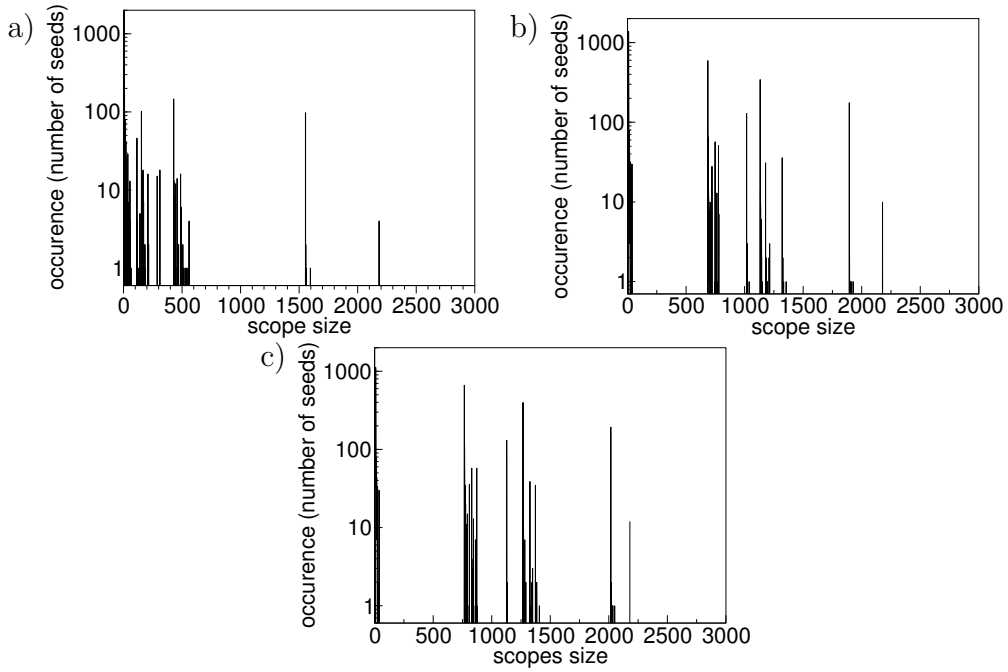


Figure 2.7: Scope size distributions of single seeds for a) the unmodified network, b) the network with the cofactors ATP/ADP, NAD^+/NADH, $NADP^+$/NADPH and conenzyme A and c) the network with computationally predicted cofactors (cf. text and figure 2.8).

factor pair must occur together in all reactions which are catalyzed by that particular cofactor pair. Clearly, such a pair should appear in a larger number of reactions in order to demonstrate that it offers a general function rather than a specialized transformation.

Searching for pairs of compounds in the KEGG network participating in a very large number of reactions yields the already analyzed cofactor pairs ATP/ADP, NAD^+/NADH and $NADP^+$/NADPH. Looking for less utilized cofactors, however, the predictions also contain often metabolized compounds like H_2O, NH_3 or acetate. This can be avoided utilizing the criterion that the two cofactors should have the same base structure. This can be enforced by comparing their structural formulas. Fortunately, the KEGG database already contains structural matchings between many reactant pairs (cf. appendix A.2).

The graph in figure 2.8 contains cofactor pair predictions where two cofactors should participate in more than 10 common reactions and where the common structure part is larger than the functional part. The pairs shown coincide with known cofactor pairs. Certain cofactors, like for example ATP,

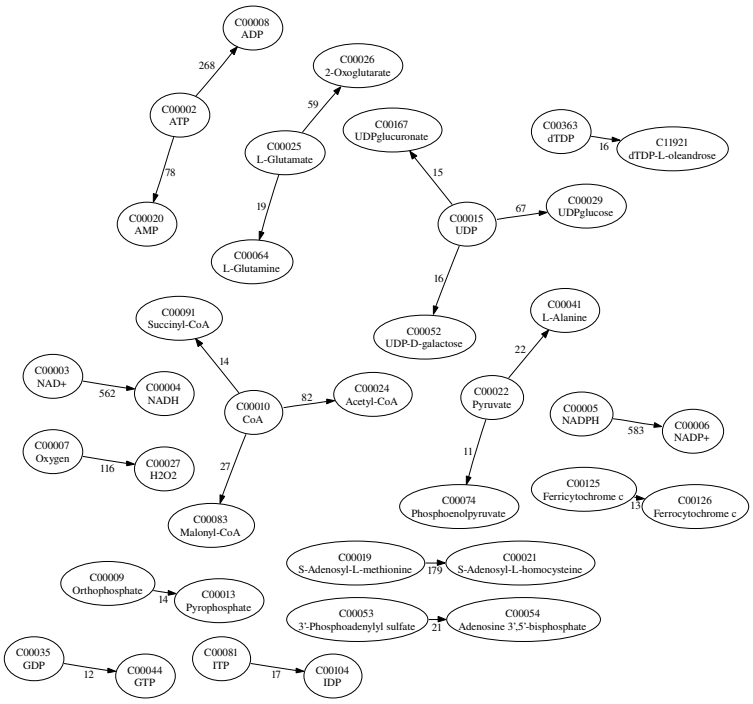


Figure 2.8: Predicted pairs of compounds which together act as cofactors. The number given next to the edges denotes the number of reactions in which the corresponding pair occurs together. Pairs were selected if they take part in more than 10 reactions and the size (number of atoms) of the functional part is not larger than their common part.

participate in more than one pair. Also coenzyme A is present and constitutes pairs with some of the possible acyl-group-coenzyme-A-composites.

For analyzing the effects of the predicted cofactor pairs on the scopes, the network is modified analogously as explained for the case of NAD^+/NADH before.

The resulting scope size distribution of all single seeds in the modified network is shown in figure 2.7c. Clearly, many scopes sizes have been increased again. However, it can also be seen that the main increase can already be explained by the first set of well known cofactors as shown in 2.7b. The reason is certainly that the first cofactors are those cofactors which participate in the most reactions and therefore also yield the most additional transformations. Furthermore, many cofactors actually offer the same functionality. For example, functions provided by the pair GTP/GDP can often already performed by the pair ATP/ADP and therefore do not increase the scopes much further.

2.6 Seed determination

So far it has been calculated which compounds can be synthesized if the network is provided with certain resources. However, it is equally interesting to know what seed compounds are necessary in order to obtain a certain set of products. One of these product sets is the complete set of compounds in the metabolic network and symbolizes the maximal synthesizing capacity. As seen in figure 2.1 a scope of a single compound covers at most 53% of all compounds of the network. Thus, more than one seed compound will be necessary. The easiest way to define a set of seed compounds which can synthesize the complete set is to take this complete set itself as seed. This trivial solution, however, does not require the network to perform any transformation at all. Therefore, in an iterative manner, compounds are subsequently removed from this seed and it is checked whether its scope still covers the complete set. If that is the case, the corresponding seed compound is removed permanently or otherwise it is returned to the seed. If all compounds are checked in the described way, one obtains a minimal set of seed compounds whose scope covers the complete network. Here, minimal means that the removal of any compound from the seed will result in a scope which does not cover the full network. The set is only locally minimal as the result depends on the order in which the compounds are checked and a different ordering may result in a even smaller set. Figure 2.9 shows a histogram of the number of seed compounds needed to cover the full network for 1000 different randomly chosen orderings. In average 535 compounds are needed. For the different runs, this number deviates only slightly from the average.

This result is somewhat surprising as it is known that many different seeds, containing different numbers of seed compounds, may result in the same scope. In section 2.3 it has been shown that for example the scope of the single seeds APS and PAPS can be reached from other sets containing more than one compound, like CO_2, H_2O, NH_3, phosphate and sulfate.

In case of a scope which requires more than 500 seed compounds one would expect a much higher variance than indicated in figure 2.9. This rises the question, whether, despite of the randomization, the predicted seed always approaches more or less the same set of seed compounds.

This is analyzed in figure 2.10. It is shown how many compounds occur as seed compounds in how many randomly calculated seeds. Interestingly most seed compounds occur in 1/2, 1/3, 1/4, etc. of the 10000 calculated seeds. A detailed analysis reveals that compounds C_i being present in half

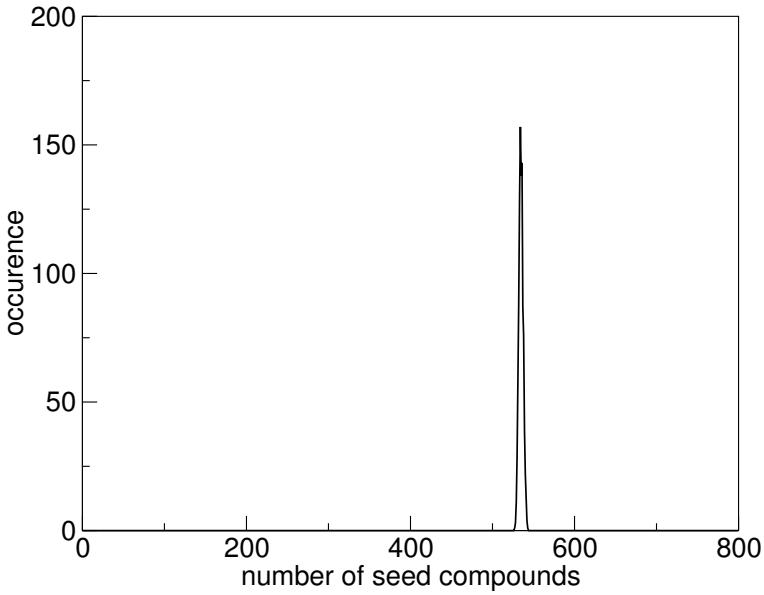


Figure 2.9: Histogram of the number of seed compounds needed to cover the full network for 1000 simulation runs. The mean of the data set is $\bar{n} = 534.7$ and its standart deviation is $\sigma = 2.6$.

of all seeds take part in the following type of reactions:

$$X_i + C_1 \rightleftharpoons Y_j + C_2 \tag{2.2}$$

where the compounds C_1 and C_2 exclusively take part in this reaction while X_i and Y_j are further connected to the remaining network. The actual number of the X_i and Y_j is not important and may even be zero. In order to have C_1 or C_2 in the covering scope, the one or the other has to be chosen as seed, resulting in an occurrence in 1/2 of all seeds for both compounds. Clearly, the topology of the above stated reaction suggests that C_1 and C_2 are a pair of interconvertible compounds under the condition that the remainder of the network is already covered, i.e. the X_i and Y_j are present. The two compounds are called exchangeable seed compounds. The exact definition will be given later in this section. In the analyzed KEGG network, there exist about 600 of such compounds, suggesting about 300 reactions of the above mentioned type.

For compounds with a frequency of 1/3 a similar argumentation can be made. Here, 3 compounds are interconvertible under the condition that the remaining network is covered. Such a situation can occur with the following

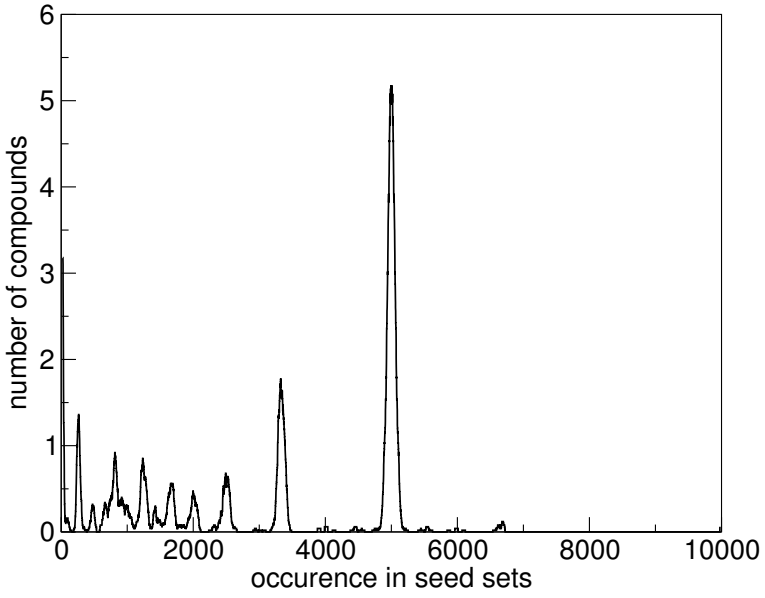


Figure 2.10: Histogram of compounds being present in a certain number of calculated seeds. The x-axis specifies, in how many seeds a compound occurred. The y-axis gives the number of compounds with that specific occurrence. 10000 seeds have been calculated. 1757 compounds occurred in at least one of these seeds. The histogram is divided into bins of size 50 and the y-axis are normalized accordingly.

topology:

$$\mathrm{X}_i + \mathrm{C}_1 \rightleftharpoons \mathrm{Y}_j + \mathrm{C}_2 \tag{2.3}$$

$$\mathrm{W}_n + \mathrm{C}_2 \rightleftharpoons \mathrm{Z}_k + \mathrm{C}_3, \tag{2.4}$$

where the C_i are otherwise isolated compounds of the 1/3 class and the X_i, Y_j, W_n and Z_k are connected to the network. Similarly the peaks for 1/4, 1/5, etc. can be explained.

Apparently, different orderings in the seed calculation algorithm result generally in different sets of seed compounds. However, the composition of such sets is strongly limited. In each random seed, exactly one of the compounds from each of the mentioned exchangeable groups must be present. As most compounds in the random seeds are compounds of that type, the variance of the total number of seed compounds is relatively low. Clearly, most of these compounds can be seen as metabolites on the edge of the network as they are connected to only a very few reactions. Thus, it can be concluded that the majority of the seed compounds is responsible for the coverage of only small parts in the surroundings of the metabolic network.

It can be assumed that these surrounding parts contain specialized pathways which may operate under certain external conditions. For example detoxification reactions may take up a toxin and convert it into a non-toxic product within a few steps. Also, there may exist less studied pathways, where certain reactions are still unknown, resulting in only loosely connected compounds.

On the other hand, the other parts of the network, which will be called the center parts, can apparently covered by only a few seed compounds as the example of APS or the small building blocks CO_2, NH_3, and H_3PO_4 demonstrates. Apparently, reactions in these parts can utilize substrates producible from a variety of common resources. Assumably, such reactions rather belong to more often utilized pathways. As here several different sets of seed compounds are exchangeable, it can be expected that their number varies more strongly.

In the following, the center of the network is examined. Therefore, the center of the network has to be well defined. Hence, a set of central metabolites is determined by considering the metabolic networks of organisms defined in the KEGG database. A set of 93 compounds, specified in table A.2 in the appendix, is present in 90% of all organisms. A seed whose scope contains this list of central metabolites is thought of covering central parts of the metabolic networks since also all intermediates necessary for the conversions between those compounds are in the scope. The above described seed calculation algorithm is applied, requiring only these target metabolites to be included in the scope.

As previously mentioned, the algorithm only finds a local minimum which is dependent on the ordering of the compounds. This is actually not a disadvantage, as there is no biological reason why a living cell should live on a global minimal set of resources. In fact, it is more useful to determine a list of several nutritional possibilities for a specified network.

Further, the ordering of the compounds can be used to prefer certain compounds over others. As the algorithm starts to remove unneeded compounds from the beginning of the list, compounds towards the end of the list are more likely to be chosen as seed compounds.

Therefore, the list of compounds was first sorted by molecular mass, putting smaller molecules to the end of the list since small metabolites make biologically more sense as nutrients than complex molecules. Compounds for which no mass was indicated in KEGG, have been given an average mass of $637Da$ (Dalton). Then, compounds defined as substrates to membrane transport processes as defined in table A.3 were moved to the end by assigning them a virtual negative mass of $-10Da$. The reason is that in living cells not all metabolites can enter the cell through its membrane and therefore

not all compound make biological sense as seed.

Still, in order to obtain several possible seeds, the ordering of the list has to be randomized. To keep the carefully prepared list roughly sorted, a special randomization is used: two randomly chosen positions in the list are exchanged with a probability

$$p = \begin{cases} \exp(-\frac{1}{\beta}\Delta m) & \text{for } \Delta m > 0 \\ 1 & \text{for } \Delta m \leq 0 \end{cases} , \tag{2.5}$$

where Δm is the difference of the molecular weights of the molecules at the two positions to be exchanged. It is positive if the heavier compound is the compound situated closer to the beginning of the list. The constant β determines the degree of disorder that is allowed in the resulting list, where a smaller value leads to less randomization. For the calculations $\beta = 20$u has been chosen and 100000 exchanges have been performed. Figure 2.11 shows the result of an example randomization.

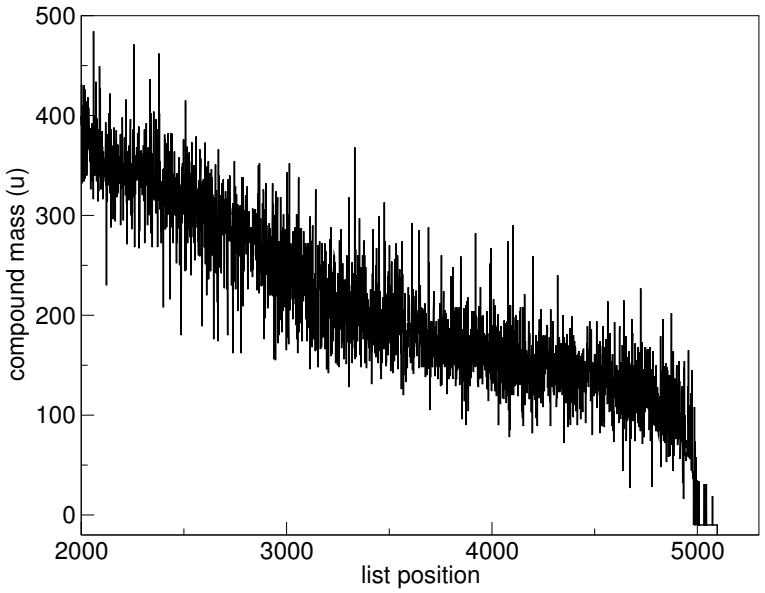


Figure 2.11: Result of a example randomization for $\beta = 20$ and 100000 position exchanges. The mass of a compound residing in a certain position is given. Only the end of the list is shown, as this is the relevant part.

Using the method described, 10000 random seeds producing at least the target set were calculated. In order to prefer substrates over products in case of irreversible reactions, the information on reversibility is included. Also, all cofactor functionalities as described in figure 2.8 are included. Among the 10000 calculated seeds, there exist 1789 distinct seeds. As the scopes now

only covers central parts of the metabolic network, in average only 4 compounds were necessary per seed (cf. figure 2.12). Altogether, 440 compounds participated in at least one of the seeds, whereas 300 of these compounds were not in the list of transported metabolites. This was expected, as the initial ordering was randomized, and also desired, as the list of transported metabolites in table A.3 may not be completely correct and consequently, other compounds should in principle be allowed in the seed. No compound was essential in the sense that it appears in all seeds.

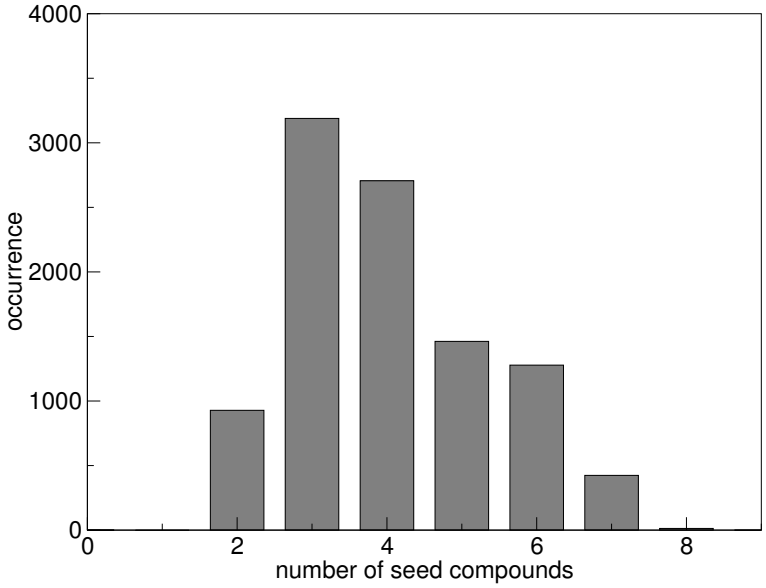


Figure 2.12: Histogram of the number of compounds in the seeds for the center metabolism. 10000 random seeds have been calculated. In average the exist $\bar{n} = 4.03$ compounds per seed. The standard deviation is $\sigma = 1.32$.

In contrast to the surroundings of the network, seed compounds of the central part cannot simply be assigned to groups, where exactly one group member has to be taken for the seed. The multitude of seeds can be more easily captured by determining which compounds can be exchanged without loosing the ability to produce the target set using the following definition:

Two compounds A and B are exchangeable if in all applicable seeds, A can be replaced by B and B can be replaced by A. A can be replaced by B, if for all minimal seeds containing A, (A,X_i), which produce the target set there exists a corresponding set (B,X_i) which also produces the target. X_i denotes the set of further seed compounds of a seed. Analogously, it can be tested whether B is replaceable by A.

As not all possible seeds (A,X_i) and (B,X_i) can be calculated, the pre-

diction of the exchangeability of A and B may be incorrect. The prediction is done in the following way: In the set of calculated seeds, all seeds of type (A,X_i) are identified. The scopes of the corresponding (B,X_i) are calculated and it is checked whether they contain the target. The compounds B are taken from the set of compounds present in at least one of the calculated seeds. The same procedure is applied for predicting the exchangeability of B by A. The two compounds are not exchangeable if there exist a X_i in the set of calculated seeds for which A and B are not exchangeable. Hence, the algorithm will not predict false negatives. However, if the two compounds are predicted exchangeable, the algorithm might have missed a X_i for which they are not exchangeable. The probability for such false positives is higher the less seeds are tested. In particular, this is the case for compounds which occur only in a few of the calculated seeds.

Figure 2.13 shows the results of this analysis as a graph. Compounds found to be exchangeable with the above definition are connected by edges. The graph consists of several clusters and isolated compounds. The compounds in the clusters are to a certain degree chemically similar. Apparently, the algorithm detects two compounds as exchangeable if they act in all tested cases as donor of the same chemical entity, like the atoms carbon, nitrogen, phosphorus and sulfur or combinations of these.

Even though the clusters can in general be categorized by their chemical content, this does not mean, that all contained compounds necessarily have a similar structure or even the same chemical elements. Figure 2.13 also indicates the chemical elements of the compounds. From there it can be seen that for example Sulfur and Carbonyl sulfide are found to be exchangeable. This means, that in all analyzed cases, Carbonyl sulfide is only used as sulfur donor. Hence, its carbon is not being used, indicating that some other seed compound can take over this role.

As described above, seed compounds responsible for covering parts of the surroundings of the network can be divided in groups, of which one compound must be in the seed. The identified clusters of seed compounds of the central part have similar meaning in the sense that only one compound of a cluster is sufficient to serve as seed compound. However, compounds of a cluster can completely replaced by compounds of other clusters possesing less or more chemical potency. In the analyzed network, some clusters provide carbon and nitrogen or carbon and phosphate, while others provide only carbon, nitrogen or phosphate. These different chemical potencies of the contained compounds is responsible for variance of the number of compounds per seed as shown in figure 2.12. The standard deviation of this number is $\sigma = 1.32$, which is about half of the standart deviation of the number of seed compounds required to cover the complete network ($\sigma = 2.6$) as shown in figure 2.9.

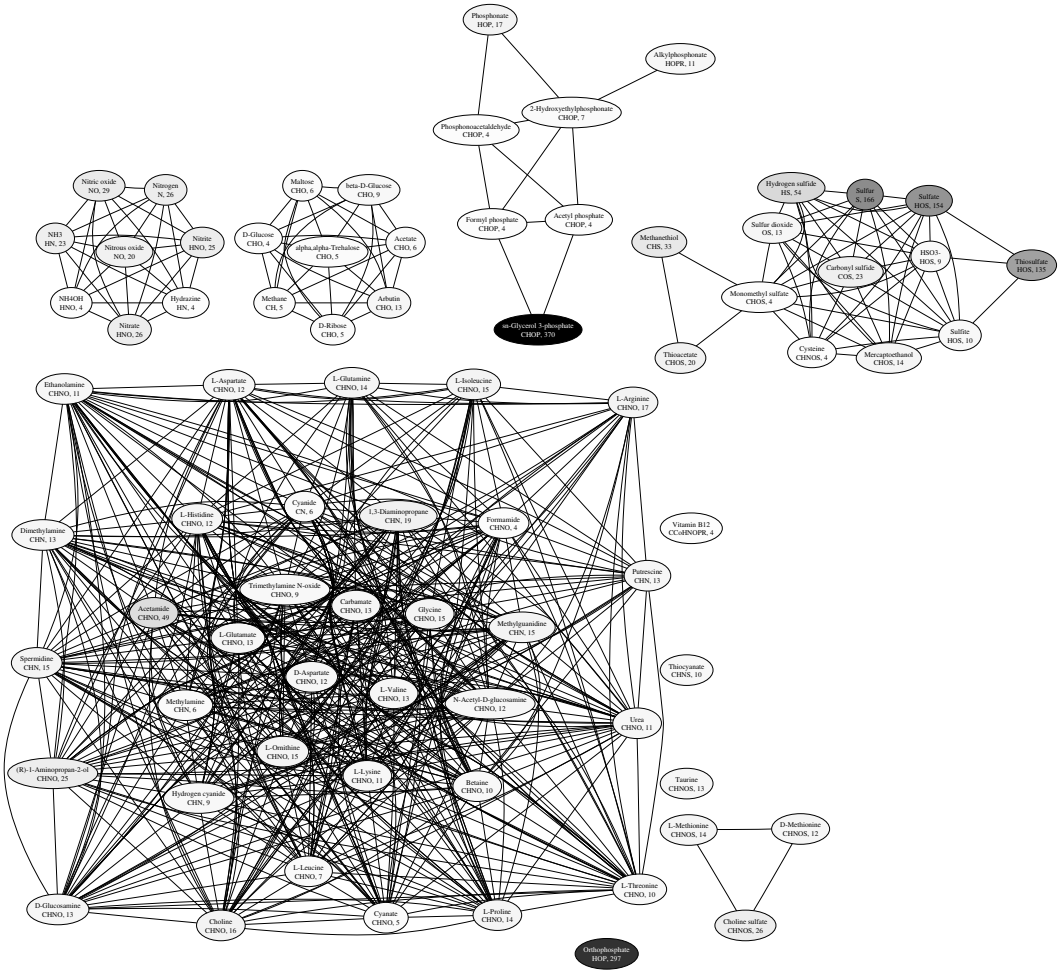


Figure 2.13: Graph representing the exchangeability of compounds in seeds. 10000 random seeds were calculated of which 1789 were distinct. In the graph, two seed compounds are connected if they are exchangeable. Seed compounds being present in less than 6 distinct seeds were removed in order to reduce the probability of false positives. The nodes give the seed compounds name, its atomic composition, and the number of distinct scopes they take part in. This occurrence is also indicated by the shading: dark for frequent compounds and bright for rare compounds.

Apparently, most of the variance in the number of seed compounds actually originates from the coverage of the center parts of the metabolism.

2.7 Distance between compounds

The expansion process can be considered as a series of consecutive synthesis steps. In each step all those compounds are synthesized which can be produced by the set of available reactions using only those compounds as substrates which were provided by previous steps. Assuming that compound B is in the scope of compound A, that is $B \in \Sigma(A)$, we define the distance $d(A, B)$ from compound A to compound B by the number of required consecutive steps to produce B exclusively from A. The distance $d(A, B)$ is not defined if B is not in the scope of A. In order to calculate that distance it is sufficient to partially expand the network starting with the seed A until compound B is reached. As the seed forms generation 1 the distance $d(A, B)$ is by one smaller than the generation in which B appears.

If A is in the scope of B and B in the scope of A, both distances $d(A, B)$ and $d(B, A)$ are defined but not necessarily the same. This asymmetry can be explained as follows. When producing B from A, a number of additional end products are generally also synthesized. The direct inversion of this process would require these end products as substrates and therefore does not represent an expansion process starting from the only seed compound B. Therefore, the synthesis of A from B in general requires different synthesis steps.

As the expansion process works as a breadth-first traversal through the network, the distance $d(A, B)$ gives the smallest number of consecutive steps in which B can be synthesized from A. However, in general more than one reaction is added per generation. Several products of in parallel attached reactions may together be required for the synthesis of the target compound B. Therefore, the number of required reactions may be larger than the number of consecutive steps.

The expansion algorithm also attaches reactions and compounds which are not required in a synthesis of compound B. In the appendix in section A.8 a method is given which can extract from the partially expanded network only those reactions which are necessary for the synthesis of B.

As examples the syntheses of citrate from pyruvate (figure 2.14a) and from pyruvate to citrate (figure 2.14b) are given. Both figures show only the reactions required for the corresponding synthesis. The production of pyruvate from citrate only requires 2 steps. The process cannot simply be inverted as acetate would be required as an additional seed. Therefore, for

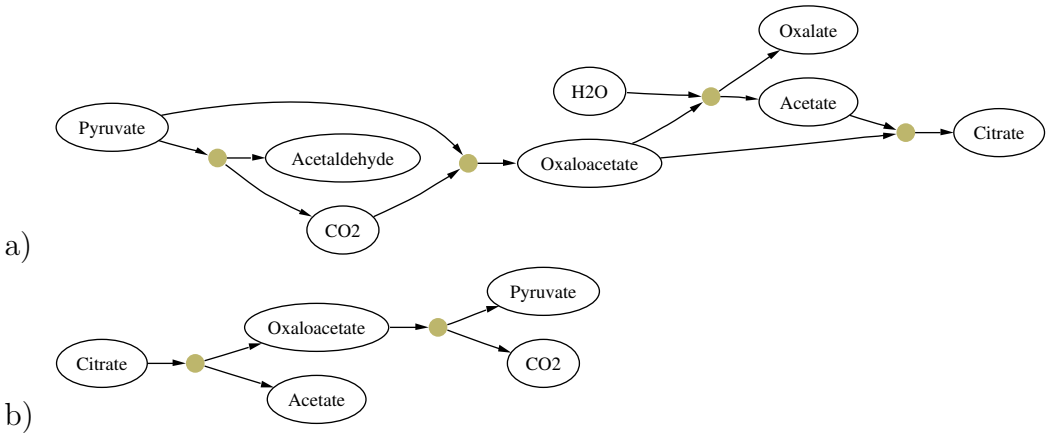


Figure 2.14: Paths for the synthesis of a) citrate from pyruvate and b) pyruvate from citrate. The first direction (a) requires 4 steps while in the other direction (b) only 2 steps are needed. Clearly, path (b) cannot simply be inverted as acetate would be required in the seed.

the reverse process, acetate has to be synthesized first from pyruvate which results in a path of length 4.

Accordingly, the distances between all compounds in the network can be calculated. Figure 2.15 shows a histogram of the distances of all pairs of compounds for which a distance exists in the way defined above. The average of these distances, $\bar{d} = 13.3$, can be seen as diameter of the network. It should however be noted that this definition is problematic as pairs of compounds which do not posses a distance do not enter this average. Thus, a very loosely connected network may still have a relatively small diameter which may appear counter intuitive.

More importantly, the distances observed here are significantly larger as reported in connection with smallworldness of metabolic networks as in Wagner and Fell [2001]. The reason for the reported small distances is mainly the fact that in their utilized graph theoretical representation (see section 1.2), many metabolites are connected through highly connected hub metabolites. For example the two compounds glucose and FAD both participate in reactions which require the hub metabolite ATP. Consequently, the two compounds are connected via ATP and have a distance of 2, even though they are chemically quite different. As the expansion process mimics the metabolic processes more accurately, its larger distances are probably more realistic.

As in the previous section, it can be assumed that the metabolic network initially possesses the functionalities of certain cofactors. Figure 2.16 shows a histogram of the distances if the functionalities of ATP/ADP, NAD^+/NADH,

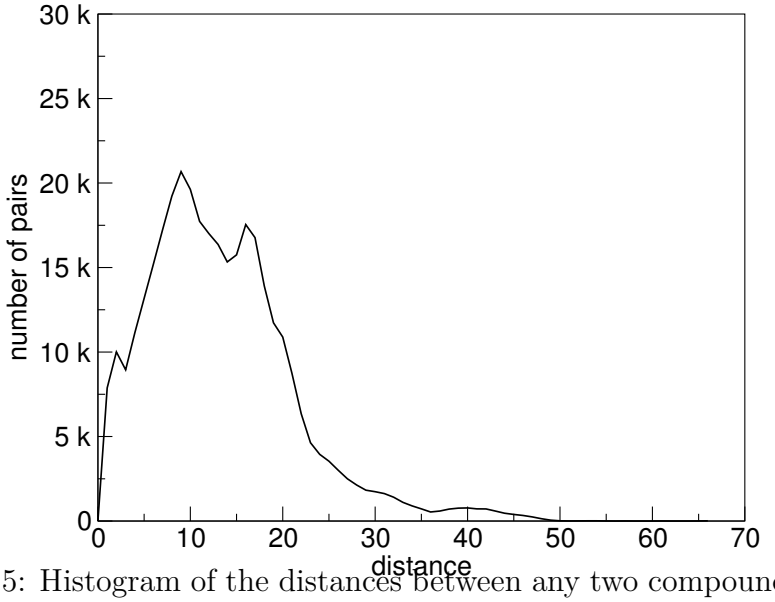


Figure 2.15: Histogram of the distances between any two compounds, where the second compound can be synthesized from the first. Pairs for which a distance is not defined did not enter the histogram.

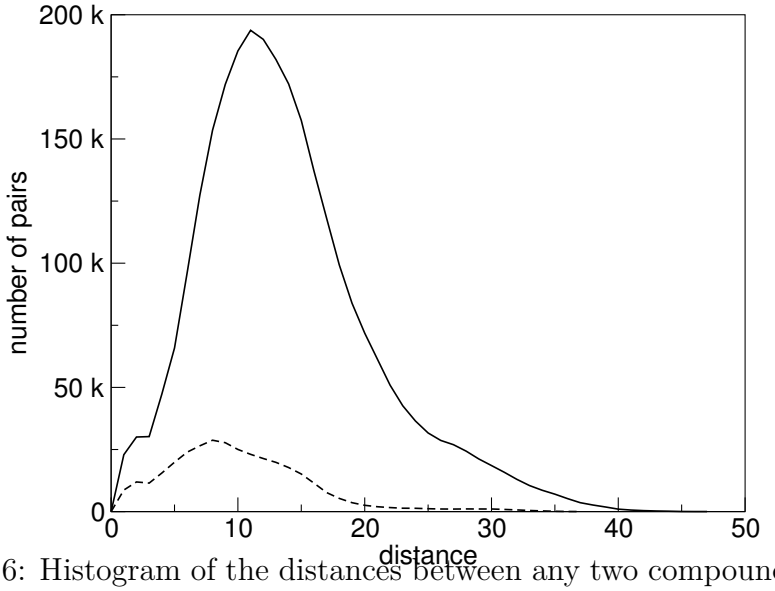


Figure 2.16: Histogram of the distances between any two compounds, where the second compound can be synthesized from the first. Here the functionality of the cofactors ATP/ADP, NAD^+/NADH, $NADP^+$/NADPH and CoA is present (solid line). The dashed line shows the distribution of distances for pairs that were connected also without cofactors present. Pairs for which a distance is not defined did not enter the histogram.

$NADP^+$/NADPH and CoA are present. Here, the average distance $\bar{d}$ is 13.9. In general, the distance of two arbitrarily chosen compounds can only become smaller if additional cofactors are present. However, the two histograms indicate that the number of connected pairs is much higher in the case with cofactors. In fact, 16% of all possible pairs are connected if the cofactors can be utuilized, whereas this ratio is only 2% if that is not the case. Consequently, the expansions reach much farther in the cofactor case leading to a higher average distance or diameter. When only considering pairs which were already connected without cofactors (dashed line in figure 2.16), the average distance is in fact smaller, namely $\bar{d} = 9.9$.

Chapter 3

Hierarchies

3.1 Relations of Scopes

In this chapter, the scopes themselves, meaning the sets of compounds that constitute the scopes, are analyzed in detail. As previously mentioned, a scope may be included in another scope. It has been shown that ATP is in the scope of APS which means that ATP can be synthesized from APS. Consequently, the scope of ATP is a subscope of the scope of APS (cf. equation. 1.17). Further, APS itself is not part of the scope of ATP, indicating that the scope of ATP is a proper subset of the scope of APS.

Generally, scopes can be related to one another by determining whether one scope is a subset of another scope. Clearly, when analyzing the scopes themselves, identical scopes of differtent seeds have to be treated as one distinct scope. Further, there exists nesting as described in equation 1.19. For distinct scopes this means that if $\Sigma_2 \subset \Sigma_1$ and $\Sigma_3 \subset \Sigma_2$ then also $\Sigma_3 \subset \Sigma_1$ holds. Clearly, the last relation does not carry any new information anymore.

Moreover, scopes may contain a common subset: $Z = \Sigma_1 \cap \Sigma_2$. It has been shown in equation 1.23 that Z itself is a scope, even though it may not be a scope of a single seed compound. Consequently, for the relations between Z, Σ_1 and Σ_2 it holds that $Z \subset \Sigma_1$ and $Z \subset \Sigma_2$.

These inclusion relations can be compiled to a directed acyclic graph, where nodes represent the scopes. Directed edges point from superscopes to their subscopes, where in case of nesting, as described above, the edge between Σ_1 and Σ_3 is omitted for removing redundancy and thereby dramatically reducing the number of edges.

Again, the analysis is performed for all single scopes in the KEGG network. For an intuitive description of the scopes they are named by one example seed compound. The resulting graph is a special representation of

the analyzed metabolic network. Unlike traditional maps, like the Boehringer map, which display adjacent reactions, this graph relates the synthesizing capacity of metabolic compounds. Through its directionality the graph implies a certain hierarchy on the scopes and thereby also on their seed compounds.

3.2 The scope hierarchy of the KEGG network

Figure 3.1 shows the hierarchy graph for the analyzed biochemical network. For a clearer layout isolated nodes are not shown. Edges always point in downward direction. Therefore nodes near the bottom of the graph represent small scopes which only have a small number of subscopes, while larger scopes are situated rather at the top. A closer investigation reveals that seed compounds of scopes in the upper part of the graph are generally more chemically complex than those belonging to scopes in the lower part. A few example metabolites are indicated in figure 3.1 to illustrate this observation.

Due to the existence of interconvertible compounds, the graph contains less nodes than possible single seeds. In particular, 4104 seed compounds yield 2922 distinct scopes. Further graph theoretical measures are given in table 3.1. These include the number of nodes (including isolated nodes), the density, the shortest path length and others. By construction, the clustering coefficient is zero as the redundant edges in case of nesting are removed. The density gives the ratio between the number of edges in the graph and the theoretical number of edges in a completely connected graph with the same number of nodes. Sources and sinks are scopes which do not have a superscope or subscope, respectively. Interestingly, only a small fraction (2.3%) of all possible source-sink pairs are actually connected by at least one path. Figure 3.2c shows the distribution of the shortest path lengths of all connected pairs.

The degree defines the number of edges pointing toward (in) or away from (out) a node. Table 3.2 shows the nodes with the largest in and out degree. Clearly, the degree of the most connected nodes differs from the average degree of 0.73. The nodes with the largest in- and especially the largest out degree are easily observable in the graph (figure 3.1). In particular, beneath the nodes of ATP and APS there exists "mushroom shaped" structures which are formed by a large number of sink nodes. Among the nodes with a high out degree there exist scopes of many cofactors. It should again be noted that always only one example seed compound is listed in table 3.2. Thus,

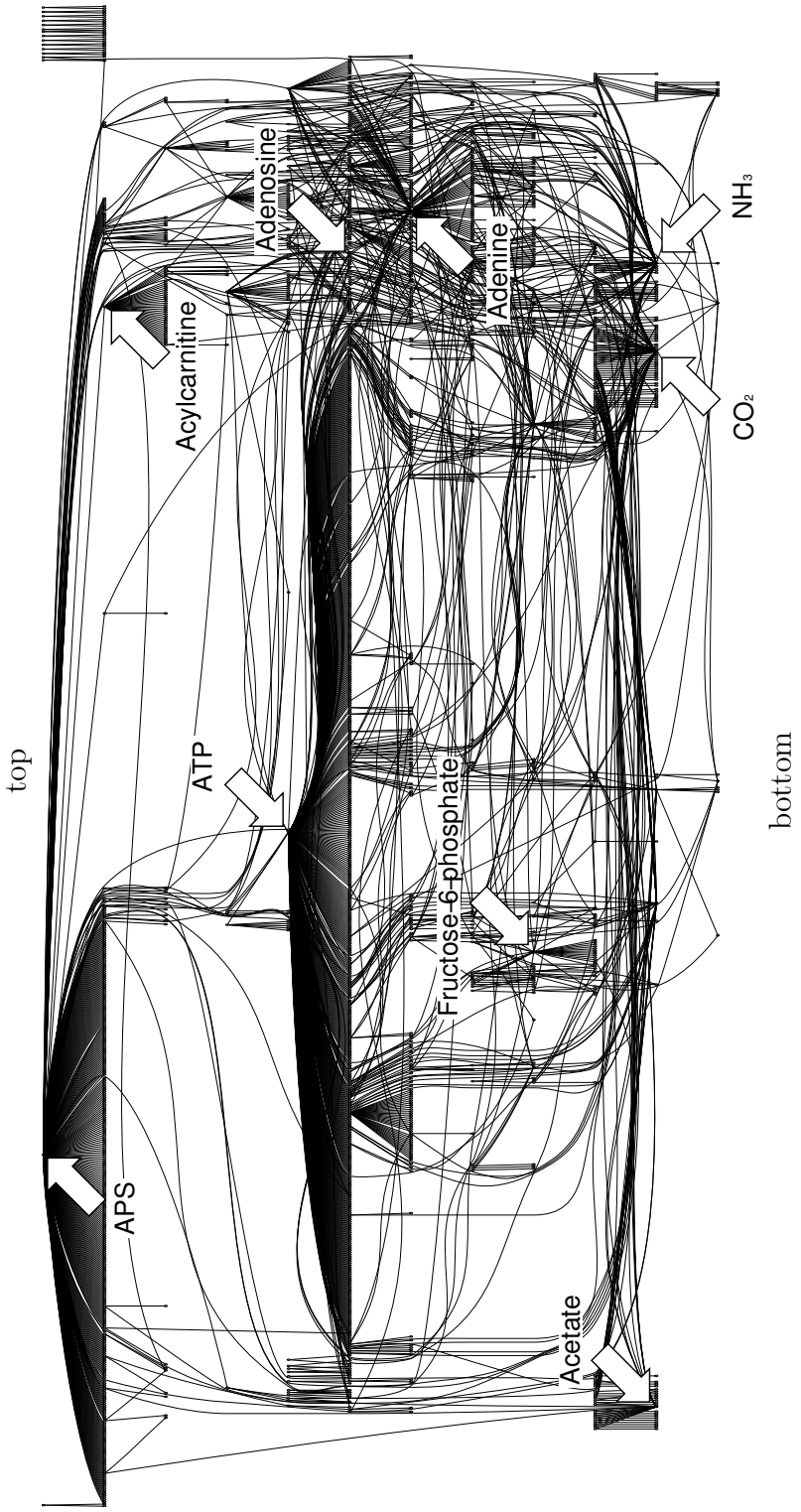


Figure 3.1: Scope hierarchy of the KEGG reference network. Nodes represent scopes and edges point from superscopes to their subscopes, from top to bottom. Well known seed compounds of a few scopes are indicated.

network property	value
nodes	2922
isolated nodes	1161
sinks	1279
sources	193
density	0.00025
mean in degree (= out degree)	0.73
shortest paths (source → sink)	5579
longest shortest path	6
longest path	11

Table 3.1: Global graph theoretical measures of the scope hierarchy as depicted in figure 3.1.

name	in degree
CO2	63
NH3	57
Orthophosphate	42
Adenine	37
Acetate	24
Pyruvate	21
D-Glucose	14
Formaldehyde	13
Sulfate	10
D-Fructose-6-phosphate	10
Hydrogen	9

name	out degree
ATP	493
APS	317
Adenine	45
Acylcarnitine	37
Isopentenyl diphosphate	33
2-Amino-5-oxocyclohex-1-enecarbonyl-CoA	26
FAD	22
Glutathione	21
D-Fructose 6-phosphate	16
Acetyl-CoA	16
Pyruvate	14

Table 3.2: Nodes of the scope hierarchy with largest in or out degree. The nodes are represented by a seed compound of the corresponding scope.

also the scopes of cofactors like NAD^+, GTP, etc. essentially possess high out degrees. As described earlier, the presence of cofactors actually activates reactions. Hence, containing a cofactor within a scope generally lets the expansion proceed considerably farther. In that way also metabolites can be reached which may be connected to only a few cofactor mediated reactions. Expansions starting from these loosely connected compounds will mostly stop very early as the the cofactors are not available. In particular for the nodes ATP and APS, this explains why so many of their successor nodes are sinks.

The distributions of the in and out degree of all nodes, as shown in figure 3.2a,b,d, indicate that besides the few nodes with very large degrees there exist many nodes with low degree. In figure 3.2b is analyzed, whether a high in degree correlates with a high out degree. The size of the circles indicate the frequency of nodes possessing a certain in degree/out degree combination. The distribution given in the figure mainly corresponds to what can be expected from a distribution of nodes with uncorrelated in and out degree. Surprisingly, the distribution of out degrees (Figure 3.2a) reveals that there exists only four nodes which have an out degree of one. An explanation is given in the appendix A.9.

When looking at the hierarchy graph in more detail, the reasons for its structure may become clearer. As mentioned above, the scopes toward the top have generally more chemically complex seeds than the ones at the bottom. For example, the scope $\Sigma(APS)$ is a superscope of $\Sigma(ATP)$ which is a superscope of $\Sigma(Adenosine)$ which again is a superscope of $\Sigma(Adenine)$. A closer look reveals that the chemical reactions which cause these inclusion relations consecutively remove the subunits sulfate, phosphate, and ribose. Furthermore, when looking at the steps between APS, ATP and Adenosine the chemical elements S and P are lost, while Adenosine and Adenine contain the same elements.

Apparently, when going from a superscope to a subscope essential parts are lost in the chemical compounds that make up the scopes. Therefore it is not possible to go back, which eventually is the reason why the graph is acyclic. These parts may be special chemical groups which can not be synthesized from the remaining parts by the available reactions, but may also be chemical elements. In the case of chemical groups the identification of the lost parts may be more difficult and is dependent on the available reactions. It can be expected that a super scope and a subscope may be merged into one scope by a modification of the metabolic network which allows for the production of the split up group from the sub scope. On the other hand, in the case of chemical elements it is inherently clear that there exist no reactions which can reproduce the lost element.

Accordingly, it can be expected that the scopes at the top of the hierarchy

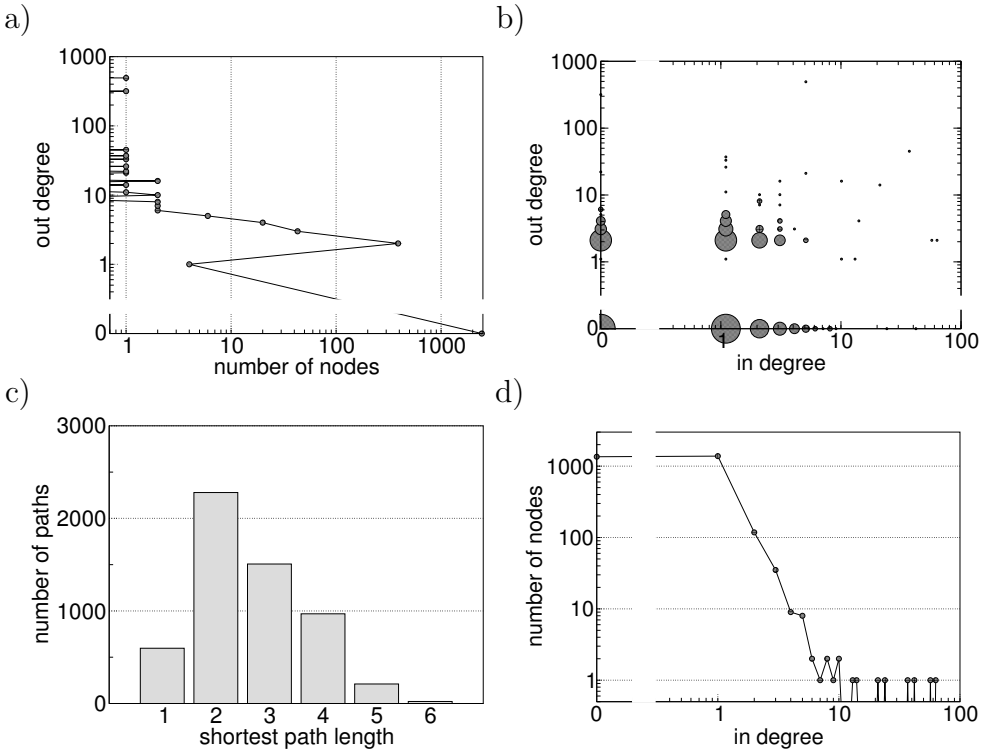


Figure 3.2: Degree distribution of the hierarchy graph. a) gives the distribution of the out degree, d) the in degree and b) the correlation of the in and out degree for the individual nodes. The size of the circles in b) denotes the logarithm of the frequency of the specific in-out degree combination. Isolated nodes are included. The figures a, b and d are arranged in a way that the the axes of adjacent figures match. Figure c) indicates the distribution of the lengths of the 5579 shortest paths from source to sink nodes, if connected.

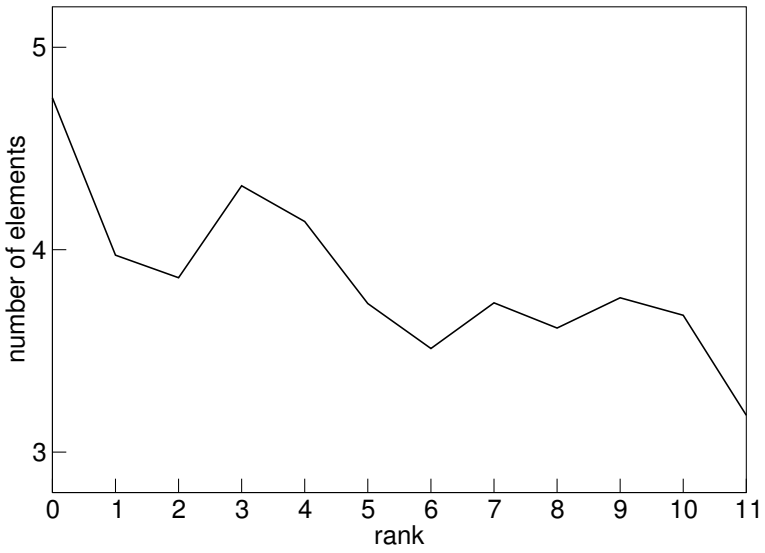


Figure 3.3: Average number of chemical elements in the compounds of the scopes representing nodes of a certain rank in the hierarchy graph. The top most nodes possess rank 0 while the lowest nodes have a rank of 11.

contain as a tendency more chemical elements than the scopes at the bottom. Figure 3.3 confirms this by indicating that the number of elements in the seed compounds generally decreases with increasing rank from the top to the bottom of the hierarchy. The rank is defined by the vertical position in the hierarchy graph. The layout algorithm (cf. Appendix A.12) positions the nodes only at discrete ranks ranging from 0 to 11. A rank of 0 is assigned to the top most nodes.

However, the dependence between the number of elements and the rank may seem weaker than expected. In fact, in some regions the number of elements increases with increasing rank. The reason is the averaging over all nodes of the same rank. For each path from source to sink, the number of chemical elements can only decrease. However, in the graph exist several parallel paths. On shorter paths the number of elements may be reduced very early which reduces the average number especially on the higher ranks. For example, the high number of sink nodes beneath the scopes of APS and ATP accounts for the low average of chemical elements at rank 1 and 2.

As mentioned earlier, not every relation of a superscope and a subscope corresponds to a loss of a chemical element. For example Adenine and Adenosine contain the same elements but yield different single scopes. In principle it should be possible to synthesize the missing ribose group of Adenosine from the remaining carbons in Adenine. This however strongly depends on

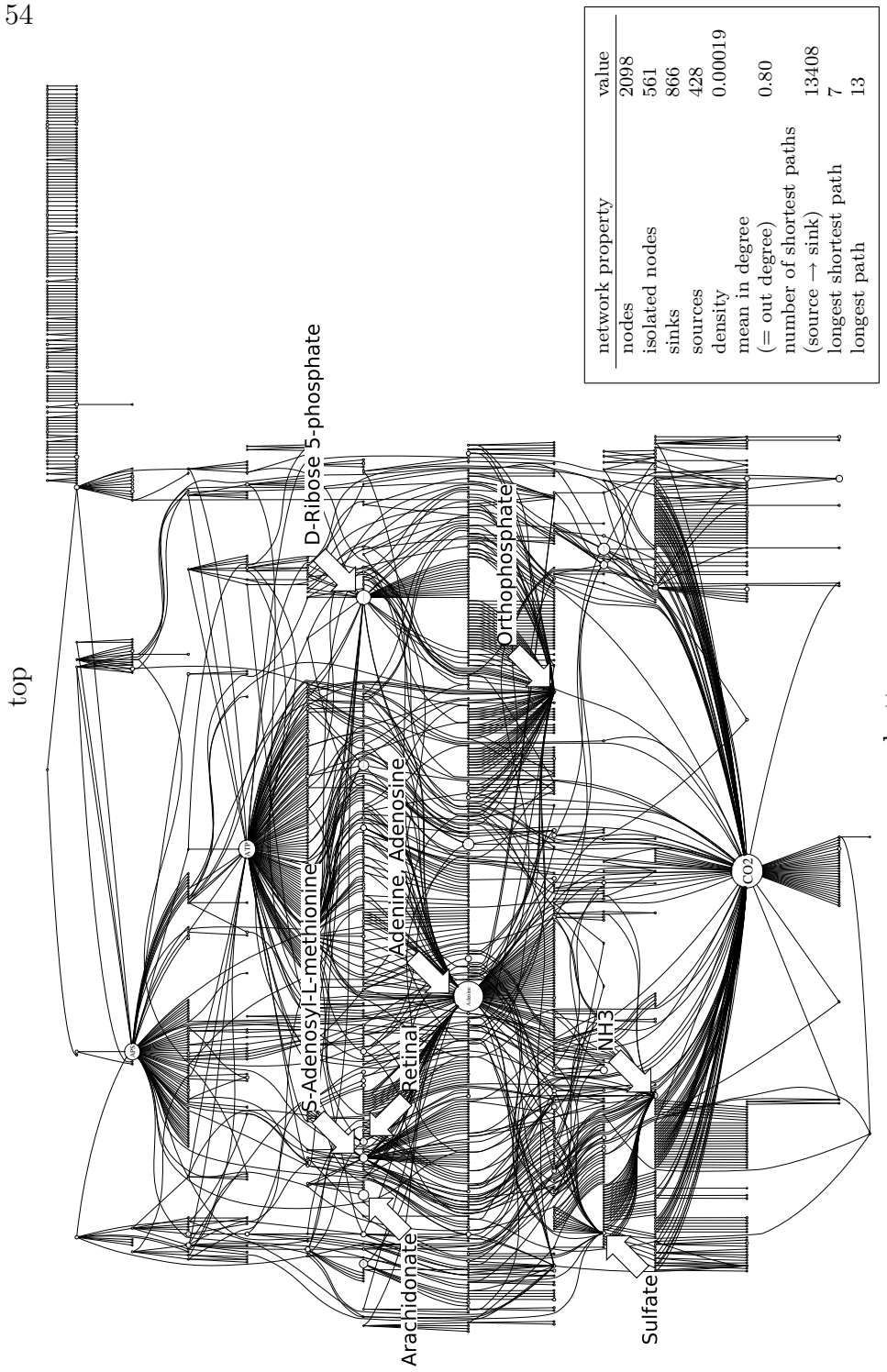


Figure 3.4: Scope hierarchy of the KEGG biochemical network including cofactor functionalities. Nodes represent scopes and edges point from superscopes to their subscopes, from top to bottom. The size of the nodes (area) roughly corresponds to the number of interconvertible seed compounds of the corresponding scope. Well known seed compounds of a few scopes are indicated. The included table contains descriptive graph theoretical measures of the displayed graph.

the available reactions. Using the current network and starting with Adenine does not lead to a production of that particular chemical group.

The situation changes when cofactor functionalities, as described in section 2.5 and figure 2.8, are considered. These functionalities change the ability of a large number of reactions to perform chemical conversions. In fact, when all cofactors are considered, Adenine and Adenosine are interconvertible. The effect on all single scopes can be studied in figure 3.4. Again, important scopes are labeled with an example seed compound. Additionally, the node size indicates the number of interconvertible seed compounds that yield the corresponding scope.

The labeled nodes generally possess a high degree and represent a large number of interconvertible seed compounds, unless they do not contain carbon. Apparently, the number of biologically relevant compounds without a carbon skeleton is low. When looking at the interconvertible seed compounds of the mentioned scopes, it becomes clear that the inclusion relations between them are not due to the presence of the chemical groups Adenine, Ribose or Phosphate, but rather indicate the possession of certain chemical elements. Most of the nodes with high degree correspond to groups of compounds containing the elements C (CO_2), N (NH_3), P (Orthophosphate) and S (Sulfate) and combinations of these CN (Adenine), CP (Ribosephosphate), CNP (ATP), CNS (S-Adenosyl-L-methionine) and CNPS (APS). Note the the elements H and O are omitted here as they are not conserved due to the presence of water in the seed.

Altogether 1187 of the 4104 compounds, hence more than 25%, are seed compounds of the these scopes. Thus, with the available reactions and cofactor functionalities, a large number of compounds can be converted into one another if they contain the same chemical elements.

On the other hand, many compounds do not fall into the above mentioned category. Some of these also form larger groups of interconvertible compounds, indicated by larger circles in figure 3.4. These often correspond to chemical groups which cannot easily be synthesized from other compounds with the same chemical elements. Two examples, Arachidonate and Retinal, are indicated in in figure 3.4. Retinal is interconvertible with many other vitamin A like compounds like Retinol, Carotene or Lutein. Arachidonate is usually synthesized from Linoleate and is a precursor for leukotrienes and prostaglandins.

It should however be noted that the existence of such distinct scopes does not necessarily infer an inability of the network to produce such a group. While Arachidonate indeed cannot be produced due to deficiencies in its synthesis pathway as described in KEGG, the situation for Retinal is different. In fact, Retinal can be produced from compounds in the CNP group (e.g.

ATP) via cartenoid and steroid biosynthesis and thus it can also be synthesized from primitive building blocks as described in section 2.3. Its synthesis is however not trivial and requires the ligation to other chemical groups. In particular, precursors of Retinal need to be phosphorylated. Therefore, starting from arbitrary compounds containing only carbon, like Retinal itself, will not lead to a production of Retinal since important intermediates are not produceable.

3.3 Modeling artificial metabolic networks

In order to analyze to which extent the conservation of chemical elements or chemical groups influences the structure of the scope hierarchy, an artificial metabolic network is utilized. Here, metabolites are composed of artificial subunits, the building blocks A, B, C, etc.. These building blocks may correspond to chemical groups or elements. Each compound may consist of 0 to N_i units of each building block i. Reactions are required to conserve the building blocks in the sense that for each type the sum of the number of units in the substrates is the same as in the products. There exists a finite number of reversible reactions transforming 1 substrate into 2 products or vice versa. In the following, only such uni-bi reactions are considered as they can already perform all possible conversions between the artificial compounds. Reactions with more substrates or products can be represented by a set of uni-bi reactions (cf. section A.10 for a proof). In a metabolic network containing all of these reactions, all compounds containing the same building blocks are interconvertible. Furthermore, compounds can be transformed into all other compounds which contain less building blocks. Such a network will in the following be called a complete artificial network. Figure 3.5 shows the scope hierarchy of such a network. Here, the compounds are composed of the building blocks A, B, and C which may be present with up to two units (N_i=2). The graph is symmetrical and shows three ranks. The only source node (top node) is the scope of the compounds containing all three building blocks. The sinks are represented by the scopes of the three single building blocks. The number of building blocks decreases with increasing rank.

The structure of this graph, however, differs significantly from the graph of the scope hierarchy of the KEGG network (cf. figures 3.1, 3.4). This is clear as also the artificial network is very different from the real network. Certainly, the real network contains a larger number of building blocks and it is not capable to make all possible conversions. In order to address these dissimilarities, a different network is created, containing a larger number of subunits, i.e. the 5 building blocks A, B, C, D, and E which can occur in

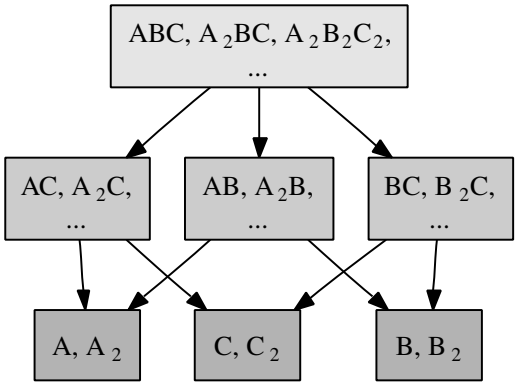


Figure 3.5: Scope hierarchy of a complete artificial network consisting of the building blocks $B_i \in \{A, B, C\}$ with N_i=2. All possible uni-bi reactions are available.

larger quantities ($N_{(A,B,C,D,E)} = (6, 4, 4, 3, 2)$). Such a network contains 2099 compounds and 186972 uni-bi reactions. The scope hierarchy of this network looks similar to the one in figure 3.5. It consists of 31 nodes and has 5 ranks.

The fact that real networks do not contain all possible reactions and compounds may have various thermodynamical and evolutionary reasons. To make up for that fact, a large number of reactions was randomly removed from the above described network, resulting in a network containing 16971 reactions. The number of reactions removed here is chosen arbitrarily. The actual dependence of the hierarchy on the number of deleted reactions is analyzed in section 4.4.

The scope hierarchy of this reduced network is depicted in figure 3.6 with descriptive properties indicated in table 3.3. The most apparent difference to the complete network is that now the scope hierarchy contains many more nodes. The reason is that due to the loss of reactions many compounds containing exactly the same building blocks are not interconvertible anymore and therefore result in different scopes. Furthermore, the graph now contains more layers which means that not in every downward step a building block is lost. On the other hand, the graph now shows features similar to the graph in figure 3.1. In particular, there exist nodes which have a large number of sink successors. Also here it can be expected that these sinks correspond to compounds connected to only a few reactions. While they still can be produced from other compounds, expansions starting exclusively from these sinks will stop early.

It is also possible to analyze the distribution of single scope sizes in the artificial network. Figure 3.7 gives these distributions for the complete network

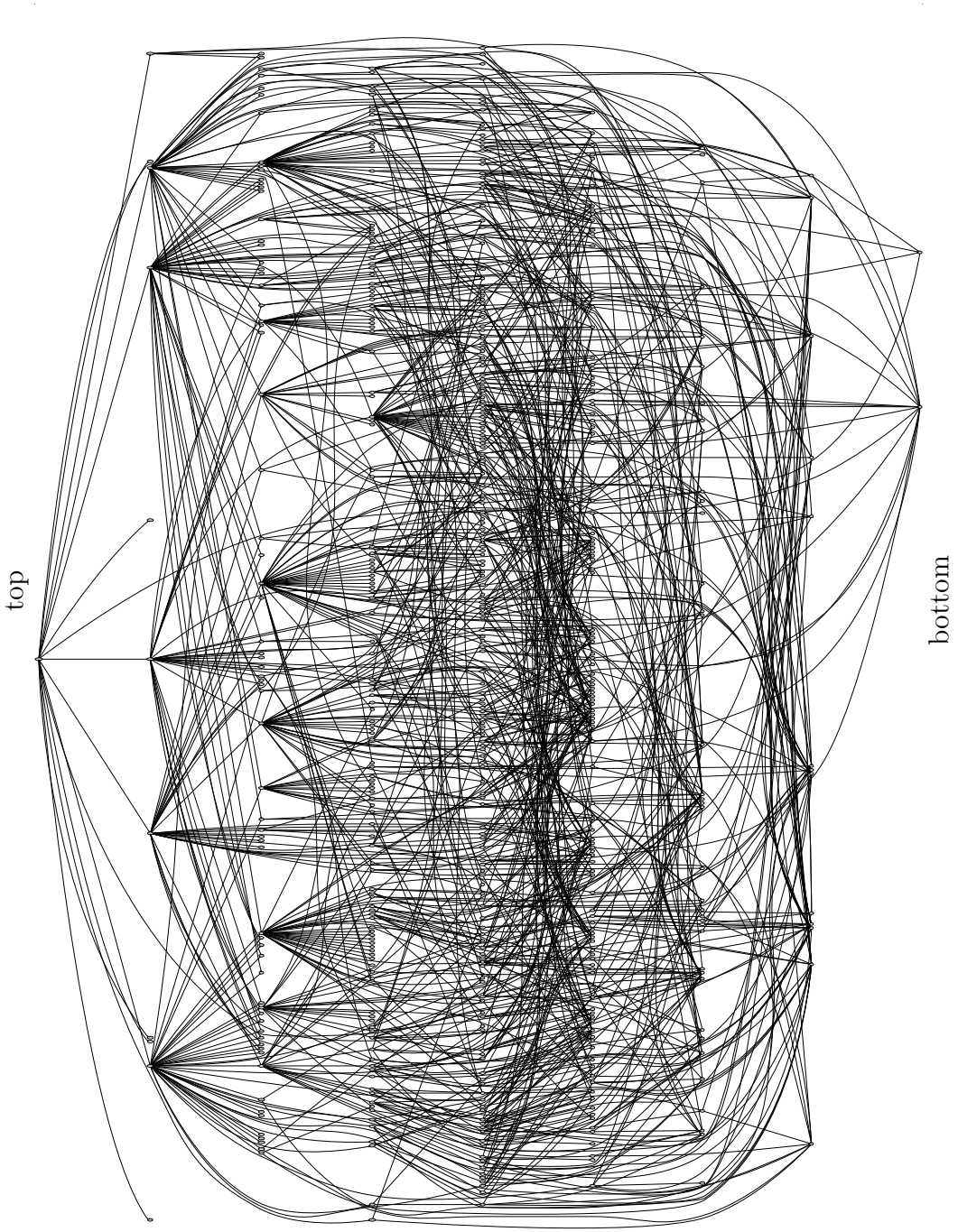


Figure 3.6: Scope hierarchy of a reduced artificial network consisting of the building blocks $\mathrm{B}_i \in \{A, B, C, D, E\}$ with $\mathrm{N}_{(A,B,C,D,E)} = (6, 4, 4, 3, 2)$. From the 186972 uni-bi reactions only 16971 reactions were randomly selected.

network property	value
nodes	525
isolated nodes	0
sinks	249
sources	1
density	0.0041
mean in degree (= out degree)	2.17
shortest paths (source → sink)	249
longest shortest path	5
longest path	8

Table 3.3: Global graph theoretical measures of the artificial scope hierarchy in Fig. 3.6.

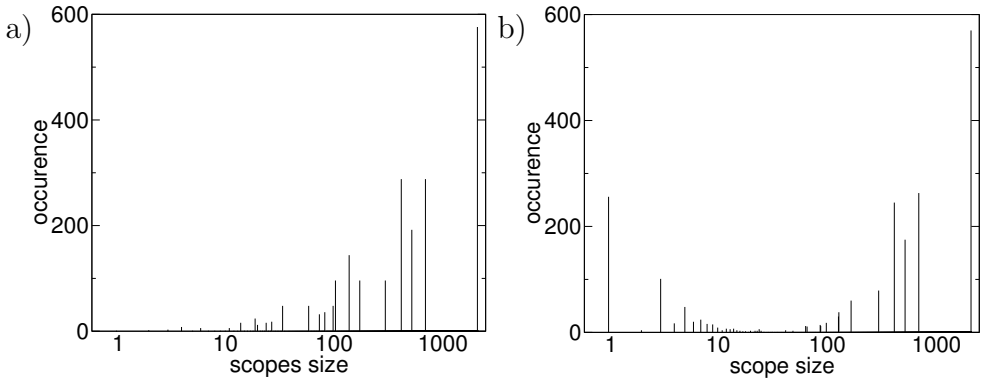


Figure 3.7: Distribution of scope sizes of single seed scopes of the artificial network with $B_i \in \{A, B, C, D, E\}$ and $N_{(A,B,C,D,E)} = (6, 4, 4, 3, 2)$: a) in the complete network and b) in its randomly reduced version as used in figure 3.6.

($N_{(A,B,C,D,E)} = (6, 4, 4, 3, 2)$) and the randomly reduced network as used in the analysis before. The complete network already shows the typical inhomogeneous distribution which has also been observed with the KEGG network. Using the building block model the large gaps especially between the larger scopes are now intuitively clear: The top element of the hierarchy represents the largest scope. Scopes of lower ranks are always significantly smaller as their are missing a complete building block and hence all compounds containing it. Scopes of the same rank possess approximately the same size depending on the specific parameters of the model.

Interestingly, also the randomly reduced network shows a similar distribution of single scope sizes, as shown in figure 3.7b. Despite the remarkable reduction in the number of reactions, the scope sizes in particular for larger scopes remained approximately the same. Only the number of seed compounds leading to a scope of a particular size is slightly reduced for larger scopes and increased for very small scopes, in particular for scopes of size 1.

Apparently, the new scopes appearing in figure 3.6 are mostly of smaller sizes. The scopes which were already present in the complete network and represent a certain building block combination remain, but are reached by a smaller number of seed compounds. These scopes will be called characteristic scopes, as they are characteristic for a certain building block set. They possess large out degrees, as many of the new scopes can still be produced from the seed compounds of the characteristic scopes, or they possess large in degrees if the corresponding building blocks can be synthesized from many of the new scopes.

Certainly, a similar argumentation can be made for the highly connected scopes in the KEGG hierarchy (figure 3.1 and 3.4). Also here, using the available reactions, many seed compounds are interconvertible if they possess the same chemical elements or groups. Analogously, these scopes are also called characteristic scopes. On the other hand, also here many compounds are only loosely connected and therefore not interconvertible with the mentioned characteristic scopes, leading to the background structure of loosely connected nodes in the hierarchy graph.

3.4 Scopes of multiple seed compounds

The hierarchy observed in the last section consists of scopes of single seed compounds. It can be assumed, however, that most scopes cannot be reached from single compound seeds but are rather the result of a seed containing several compounds. The KEGG network as used here contains 4104 compounds which allows for 2^{4104} different sets of compounds. Each set can be

used as seed. Each set of seed compounds will expand to a scope, which itself is also a member of all possible sets of compounds. As observed with single scopes, different seeds may converge to the same scope. It can be expected that the total number of scopes will be larger than the number of single scopes, i.e. 2923, and smaller than the total number possible sets of compounds, i.e. 2^{4104}. These considerations suggest that scopes of multiple seed compounds play an important role and may be worth being analyzed in more detail.

To begin with, scopes of two seed compounds are analyzed. As interconvertible compounds always behave the same in the scope analysis it is sufficient to construct the seed compound pairs from a reduced list of compounds where groups of interconvertible compounds are represented by only one member. Thus, a total of 4270503 pairs of seed compounds (as there exist 2923 of such groups) exist. Again, water is assumed to be present during the expansion processes.

The scope calculations revealed that most of the resulting double scopes are simply set unions of the corresponding single scopes. In fact, this is the case for the vast majority, for 4149610 pairs. This observation can be explained as follows: As seen before, most single scopes are rather small. Therefore, it is quite likely that two such scopes are situated in different network regions so that they are not adjacent to a common reaction. Consequently, the union of these single scopes cannot be expanded any further. The resulting scopes are mostly unique. See section A.13 for details.

Only 120893 pairs possess scopes which are larger than the union of their single scopes. Scopes of different seeds may coincide. In fact, the 120893 seeds result in only 62341 unique scopes. Still, most of the seeds yield unique scopes or are interconvertible with only a few others. A smaller number of seeds forms large groups of interconvertible seeds. Table 3.4 gives an overview of the ten largest groups of interconvertible seeds. Also, the distribution of the sizes (cardinality) of all groups is shown.

It is interesting to note that the scopes belonging to the largest groups of interconvertible seeds are actually identical to well know single scopes. The largest group, representing 2936 interconvertible seeds, yields a scope identical to the scope of ATP. 1346 seeds have the same scope as APS. Groups which do not correspond to single scopes can be identified as characteristic scopes of particular chemical groups and elements. The chemical groups shown have been found in each seed of the corresponding groups.

Interestingly, there exist 2 distinguishable scopes which are characteristic for the elements C,N and S. A closer investigation reveals that the seeds of the larger scope contain at least one sulfur atom with 6 covalent bonds, while in the smaller scope, seeds contain sulfur with a maximum of 4 bonds. Appar-

group size	scope size	single scope	composition
2936	1554	ATP	CNP
1346	2183	APS	CNPS
532	428	Adenine	CN
405	525	-	CN,CoA
241	506	Taurocyamine	CNS^6
208	1563	-	CNP,Cholate
201	2218	-	CNPS,Riboflavin
199	1568	-	CNP,Hexadecanal
169	1574	-	CNP,Thiazole
150	487	Glutathione	CNS

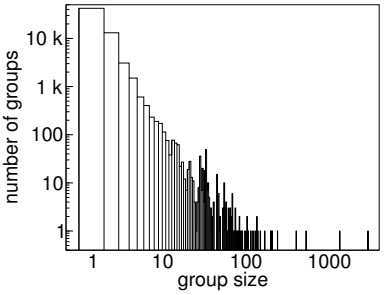


Table 3.4: Groups of interconvertible double seeds. The ten largest groups of interconvertible seeds are shown along with the corresponding scope sizes, an identical single scope, if any, and the chemical composition of the seed. The composition comprises the chemical elements and chemical groups present in the seeds. The presence of a chemical group means that this group or a chemically related group is present in each seed of the corresponding group of interconvertible seeds. Also, a distribution of the sizes of all groups is given, indicating how many groups have a certain number of members (group size). 42186 seeds are part of groups with only 1 member which means that they are not interconvertible with any of the other seeds.

ently, if cofactor functionalities are not present, it is not possible to produce a hexavalent sulfur from tetravalent sulfur. If the cofactor functionalities are present, these two groups of seeds become interconvertible.

Figure 3.8 shows the distribution of scope sizes of all 4270503 double scopes. The figure shows a stacked graph indicating the number of expanding scopes (gray - scopes larger than the union of the corresponding single scopes) and non-expanding scopes (black - scopes being the union of the corresponding single scopes) with a given scope size.

Analogously to the distribution of single scopes sizes, also the majority of double scopes seems to be small. Further, the distribution is also not homogeneous but rather consists of separate bands. Interestingly, double scopes seem to integrate into the structure formed by the single scopes. In fact most of the observed bands of double scopes are situated next to characteristic single scopes.

A similar picture can be observed when looking at scopes of more than 2 seed compounds. Certainly, it is not possible to analyze systematically all possible combinations of seed compounds. Therefore, in figure 3.9, the distribution of scope sizes of 10000 randomly chosen seeds is shown. The

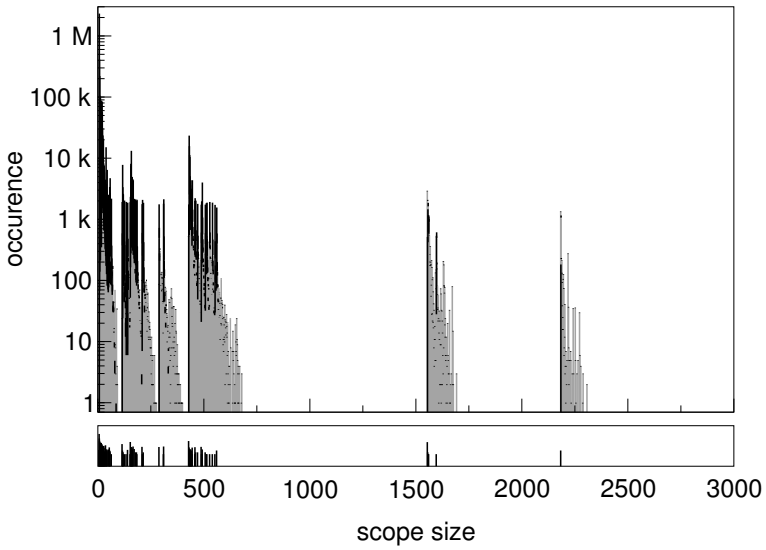


Figure 3.8: Distribution of scope sizes of double scopes. The graph is stacked. The lower gray part represents the expanding double scopes while the black top part corresponds to non-expanding scopes. In the histogram, for each seed one scope is counted, i.e. seeds converging to the same scope are counted separately. Due to the logarithmic scale, the number of non-expanding scopes may appear much smaller than it actually is. For comparison, the sizes of the single scopes are indicated in the small bar-like histogram at the bottom.

seeds where chosen to contain in average 25 random compounds. Also here, the scopes form separate bands adjacent to characteristic single scopes.

A more detailed analysis shows that many of the random multi scopes next to the characteristic single scopes in fact include these single scopes. Figure 3.9 includes a curve depicting the size of the common subset of all scopes larger than a certain size. It turns out that all scopes larger than 1554 contain a common subset of size 1554 (i.e. the scope of ATP) and all scopes larger than 2183 have the scope of APS (size 2183) in common.

Eventually it has been analyzed, how the distribution of scopes sizes is influenced by the number of compounds in the seed. In figure 3.10 a two-dimensional distribution is shown, displaying the scope size distributions in dependence of the number of seed compounds. 10000 random scopes were calculated for each analyzed number of seed compounds. It can be seen that for larger numbers of seed compounds (i.e. larger than about 50) the formerly observed bands disappear and a single band remains. The sizes of the scopes in this band monotonously increase with increasing number of seed compounds. Consequently, this increase culminates in the case where

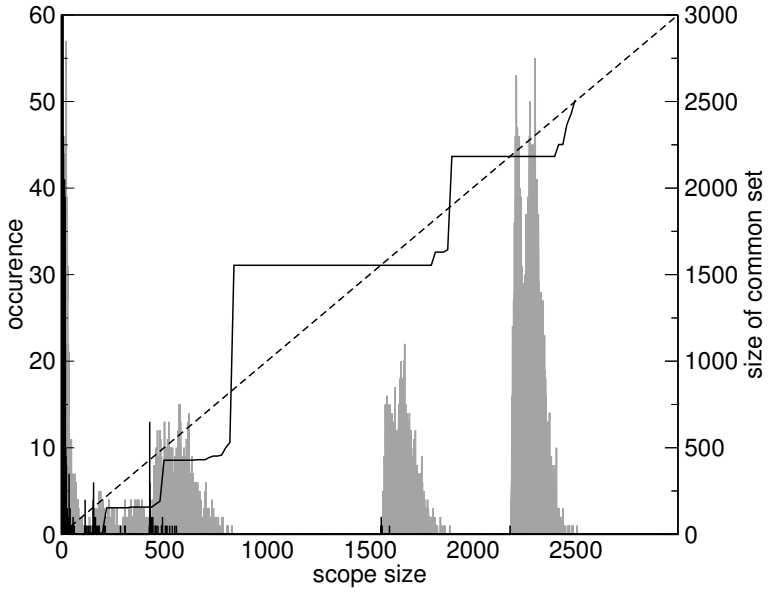


Figure 3.9: Distribution of scope sizes of 10000 scopes of in average 25 random seed compounds (gray) and of single scopes (black). The solid line gives the size of the common subset of all scopes larger than a certain size. Whenever this line intersects with the bi-secting line (dashed) vertically from left to right, a scope at that size exactly coincides with the cut set of all larger scopes. This is the case for the scope of ATP and APS which are hence contained in all calculated scopes larger than themselves.

all compounds are in the seed and hence also in the scope.

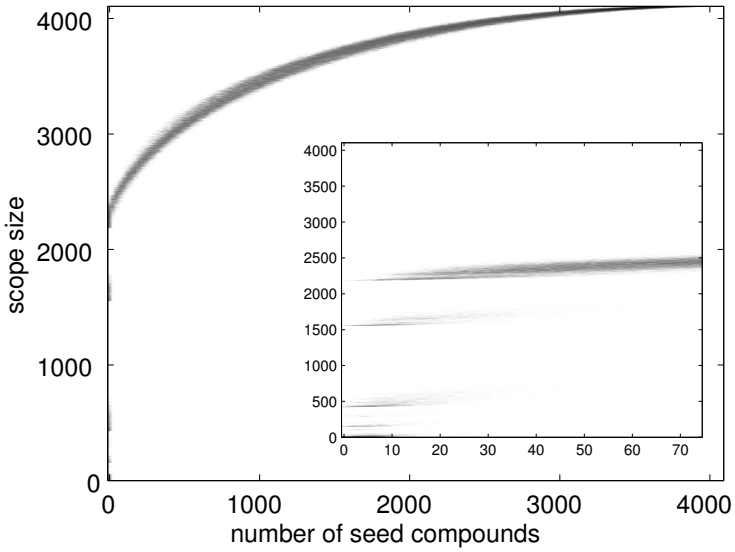


Figure 3.10: Distributions of scope sizes in dependence of the number of seed compounds. The distribution at x=1 represents the distribution of single scopes as in figure 2.1, x=2 corresponds to figure 3.8 and x=25 to figure 3.9. For x=4105 (i.e. the total number of compounds) all compounds are seeds and the complete network is covered. The inset gives a magnification of the distributions for smaller numbers of seed compounds.

The distributions observed for the scopes of multiple seed compounds can be explained to a large extent by the hierarchy of the single scopes. At least larger multi scopes are generally arranged in bands on the right side (toward larger scope sizes) of characteristic single scopes. Gaps exist in particular between the bands of larger scopes. The reason is the large size of the two characteristic scopes, the scope of ATP and APS, in comparison to the relatively small size of all other single scopes. Apparently, a multiscope is in one of the upper bands if it contains the scope of ATP or APS. If it does not contain one of the two scopes it is relatively small.

For an increasing number of seed compounds it becomes more and more likely that the scope of APS is included. Therefore, for large seeds only the largest band remains. The scope sizes in this band further increase with increasing number of seed compounds indicating that in general each new seed compounds adds a few new compounds to the scope.

3.5 Multi scopes in artificial networks

To understand the behavior of multi scopes it is again useful to analyze them in an artificial network as described by the building block model in section 3.3. Analogously to the scopes in the KEGG network, multi scopes in artificial networks also show certain structuring. Figure 3.11 shows the distribution of the scope sizes of 10000 randomly selected sets of 3 seed compounds in a network with $N_{(A,B,C,D,E,F)} = (7, 2, 2, 1, 1, 1)$. Again, the multi scopes tend to cluster next to characteristic single scopes.

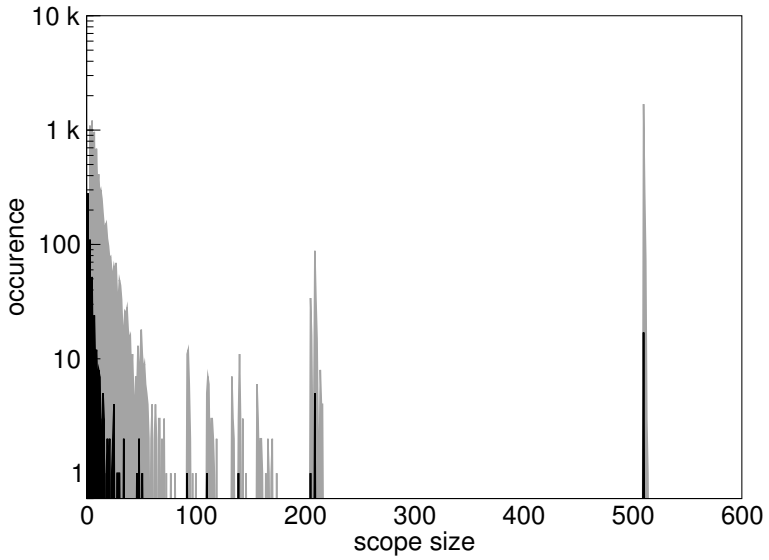


Figure 3.11: Distribution of scope sizes of 10000 scopes of 3 random seed compounds (gray curve) in an artificial network with $N_{(A,B,C,D,E,F)} = (7, 2, 2, 1, 1, 1)$. The network contains 575 compounds and 626 reactions (of 16928 possible reactions). The black curve indicates the single scope distribution.

It should be noted that this clustering of multiscopes is best observed if the building blocks are unevenly distributed like for $N_{(A,B,C,D,E,F)}$=(7,2,2,1,1,1). Only with such an distribution there exists a few very large scopes along with a vast number of very small scopes. Only then, some multi scopes contain these large scopes together with a few additional compounds which lead to the observed bands in the scope size distribution.

The situation in the KEGG network is actually similar. Here, the number of C atoms can be very large, while other atoms participate only in smaller amounts. The H atoms can be ignored in this case. Indeed, they do participate in large numbers, but they are coupled to the C atoms and cannot exist

in large numbers with out a large C-skeleton in the background.

3.6 The total number of scopes

It is clear that in an artificial network containing all possible reactions, every multi scope will coincide with a single scope. In fact, it will be identical with the single scope that contains exactly the same building blocks.

When removing reactions from the network, interconvertibilities will be lost. As described in section 3.3 this will lead to a larger number of single scopes since formerly interconvertible compounds now posses different scopes. Furthermore, also scopes of more than one seed compound may now not be interconvertible anymore with their corresponding single scopes and therefore yield their own scopes.

While in the case of single scopes their total number is limited to the total number of compounds, the number of multi scopes may become dramatically large when more and more reactions are removed from the network. As described before, the number of scopes may become as large as 2 to the power of the number of compounds. This would be the case if no seed is interconvertible with any other seed which can be reached by removing all reactions from the network.

For a small network with four building blocks and $\mathrm{N}_{(A,B,C,D)} = (2, 1, 1, 1)$, possessing 23 compounds and up to 58 reactions, the total number of scopes in dependence of the number of reactions removed has been calculated and depicted in figure 3.12. As for the generation of this graph the reactions have to be removed consecutively, the actual result depends on the order of the removal. However, runs following different orders have shown similar results. The most interesting fact of this graph is that the number of scopes stays relatively low when the first reactions are removed and only increases strongly during the removal of the last reactions.

Furthermore, it turns out that for many of the late removals, the number of scopes doubles. This can be explained by looking at the last reaction in detail: The last reaction reads $C_1 \longrightarrow C_2 + C_3$ (without loss of generality). Without that reaction the three compounds would allow $2^3 = 8$ possible scopes. With that reaction there exists only 4 scopes, namely no compound, C_2, C_3 and all three compounds. The consideration of the scope containing no compound is important as scopes defined by these three compounds combine with the scopes defined by the remaining network. Hence, the presence of the last reaction halves the total number of scopes in an otherwise empty network.

It is clear, that further reactions, not being connected to other reactions

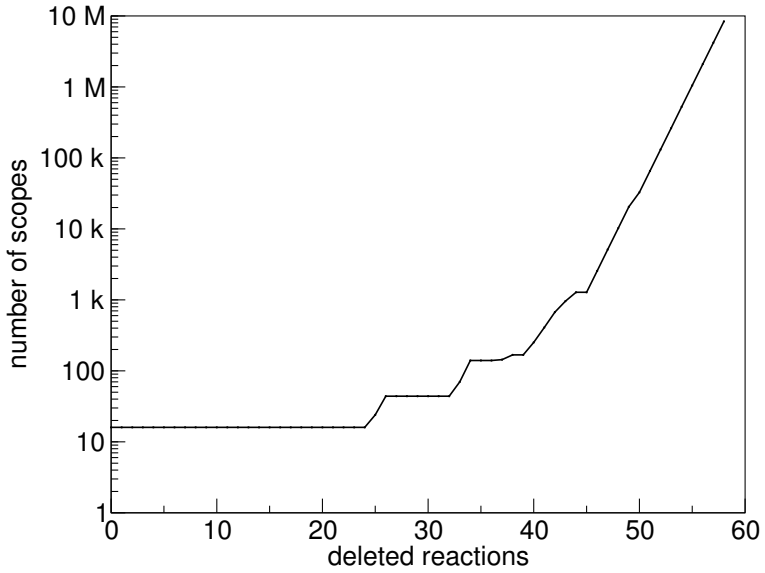


Figure 3.12: Total number of scopes in an artificial network with $N_{(A,B,C,D)} = (2, 1, 1, 1)$ and a maximum of 58 possible reactions. The curve shows the total number of scopes in dependence of the number of reactions randomly and consecutively deleted from the network.

will independently half the number of scopes. This explains the behavior at the end of the curve in figure 3.12.

As discussed before, the total number of scopes in the KEGG network is between 4211951, the number of double and single scopes, and 2^{4104} the number of all possible sets of compounds. It can be assumed that the number of scopes is much larger than the given lower bound. Many of these scopes are certainly just set unions of other existing scopes.

The upper limit of the number of scopes can be reduced by determining which fraction of the total number of compound sets are actually scopes. This is done by testing random compound sets whether they are scopes or not. Running this test for a couple of hours yielded 16000000 checked sets of which none was a scope. This essentially reduces the upper limit by a factor of 2^{24} to still 2^{4080}. Apparently, the number of scopes is much smaller than the upper limit and therefore the determination of the fraction is computationally too expensive. A better upper limit can be estimated by considering the formerly observed fact that a single reaction already dramatically reduces the number of total scopes. Generally, a $n \leftrightarrow m$ reaction connected to otherwise unconnected compounds reduces the number of scopes by a factor

$$r = 1 - \frac{1}{2^m} - \frac{1}{2^n} + \frac{1}{2^{n+m-1}}, \tag{3.1}$$

as explained in detail in the appendix A.14. Adding further reactions to a network can never increase the number of scopes as an additional reaction does not destroy any interconvertibilities provided by other reactions. In the appendix it has been shown that 339 reactions can be included in the empty KEGG network in a way that each compound is connected to only one reaction. Then, the upper limit can be reduced by a factor of about 2^{320}.

Hence, even though the total number of scopes could not be determined for the KEGG network, it could be confined to values between 4211951 and 2^{3784}.

3.7 Hierarchies of multi scopes

Analogously to single scopes, also all multi scopes in the KEGG network form a scope hierarchy. As this hierarchy graph is far too large for being calculated or even displayed, only aspects of the graph can be discussed. As pointed out for single scopes in artificial networks, there exist characteristic scopes, which essentially represent a certain set of building blocks. As observed, multi scopes often coincide with such characteristic scopes or cluster next to them. This suggests that the multi scopes contain the building blocks specific to the neighboring characteristic scope. Further optional building blocks in the multi scopes apparently have a smaller impact on the scopes size, leading to a slightly larger scope size.

The inclusion of the characteristic scopes can be confirmed in the scope hierarchy. There, characteristic scopes are expected to be sub scopes of the corresponding multi scopes. Figure 3.13 shows a part of the overall scope hierarchy, namely the hierarchy of characteristic single scopes extended by 250 multi scopes, each generated from 5 random seed compounds.

This hierarchy confirms that the characteristic scopes are actually subsets of the corresponding multi scopes. The fact that most of the multi scopes are directly and exclusively connected to the corresponding single scopes emphasizes the special role of these scopes in metabolism.

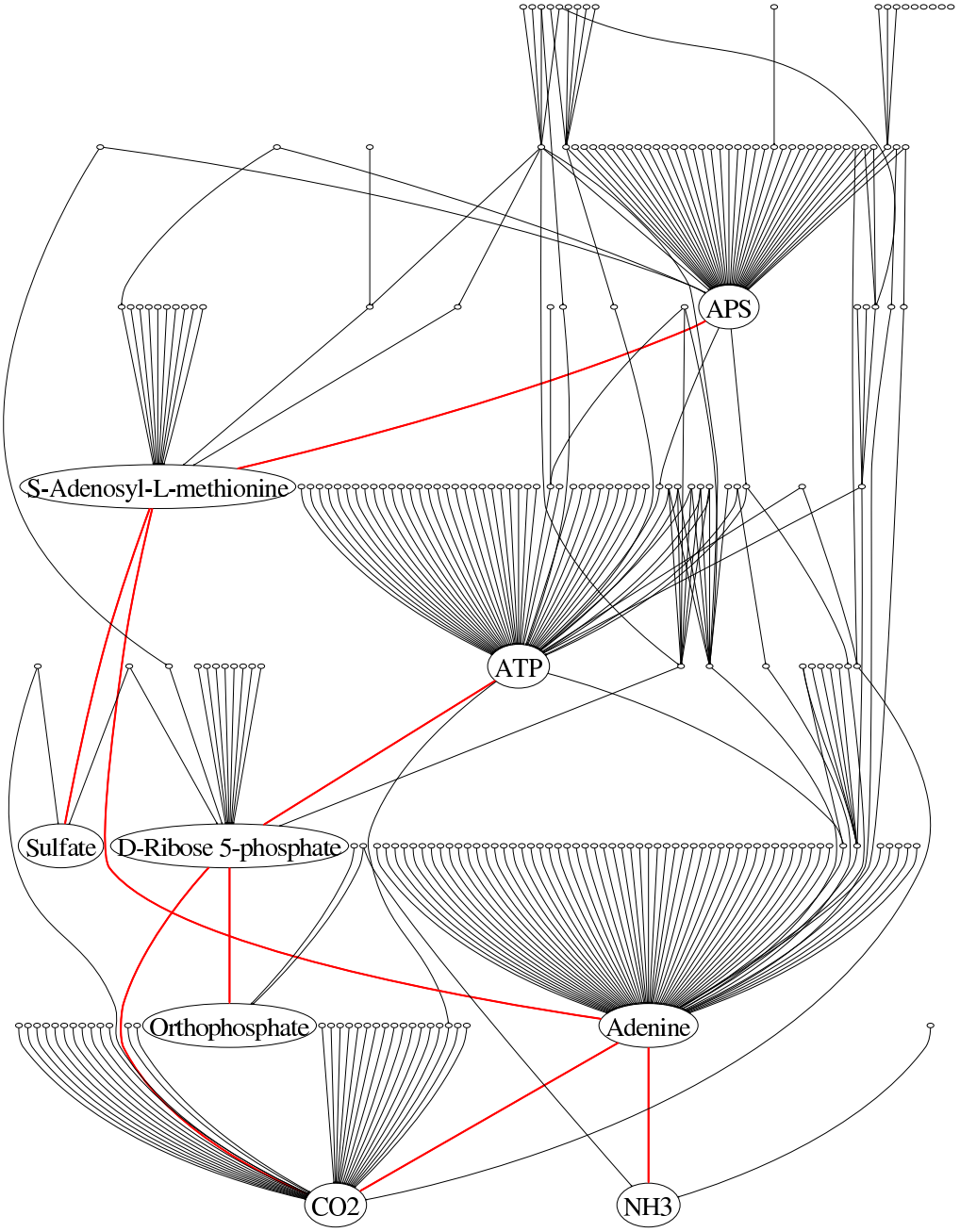


Figure 3.13: Scope hierarchy of multi scopes. 250 random multi scopes with 5 random seed compounds have been calculated and put together with selected characteristic scope in one hierarchy graph. For the calculations cofactor functionality has been assumed.

Chapter 4

Variation of the underlying network

This chapter is dedicated to the question how the so far obtained results are influenced by changes in the underlying network.

The analysis of the effect of network modifications is useful for several reasons. First, a more technical reason is that biological networks available today only represent the current knowledge on the real situation. It is therefore necessary to assess how future improvements of the network data will affect the results of the scope analysis.

Second, it is useful to analyze the robustness of such networks. Organisms and thereby their metabolisms often suffer mutations which may inactivate certain reactions. It is possible to analyze the effect of such mutations on the synthesizing capacity.

Third, metabolic networks do look different in different organisms. Even though several pathways follow similar objectives in these organisms, it can be assumed that the differences in the networks will also show differences in their synthesizing capacities.

Ultimately, metabolic networks are the result of an evolutionary process. This process inherently modifies the network topology in order to adapt it to certain external forces. Analyzing the effect of network modifications on the scopes will possibly uncover hidden principles of this evolution.

4.1 Properties of scopes on variable networks

As mentioned, in this chapter the behavior of scopes is analyzed if the underlying metabolic network is changed. To reflect the dependency on the

network, the scope operator is modified accordingly:

$$\Sigma^{\Re}(S) \tag{4.1}$$

where $\Re$ is the set of reactions in the network and S is the set of seed compounds.

Obviously, additional reactions cannot lead to a reduction of the scope. Formally,

$$\Re_1 \supset \Re_2 \Rightarrow \Sigma^{\Re_1}(S) \supseteq \Sigma^{\Re_2}(S) \tag{4.2}$$

This means that in general larger networks produce larger scopes. Small changes in the network structure may have a large effect on the scopes, as will be demonstrated in this chapter.

For scopes of the same seed compounds on different reaction sets holds

$$\Sigma^{\Re_1 \cap \Re_2}(S) \supseteq \Sigma^{\Re_1}(S) \cap \Sigma^{\Re_2}(S), \tag{4.3}$$

as there may be reactions in one of the sets which can use compounds produced by the other set.

It may be useful to define an associated set of reactions which contains those reactions used to produce the scope from the seed. It is defined as the set of reactions whose substrates and products are completely in the scope. Formally, it can be represented by the symbol

$$W^{\Re}(S). \tag{4.4}$$

It should be noted that each scope Σ an associated set of reactions W is uniquely assigned.

4.2 Robustness against single deletions

From equation 4.2 it follows that if reactions are added to a network, the scope sizes may increase or remain the same. In turn, it can be concluded that a removal of reactions may decrease the scope size or leaves the scope unchanged. Technically, the analysis of the reduction and addition of reactions is the same as in both cases there exist two sets of reactions, of which one is included within the other. When analyzing the KEGG network it is difficult to define rules for addition of new arbitrary reactions. Therefore, in the following only the reduction is used for the analysis.

In this section it is analyzed to what extent the removal of single reactions affects the scopes. The calculations were done on the KEGG network. For each analyzed scope it is determined, how the scope size is reduced if

each reaction from the associated set of reactions W is separately removed. Clearly, the deletion of reactions not included in W cannot have an effect.

In figure 4.1, the resulting effects are depicted for the scopes of ATP, APS, L-Glutamate and for the multi scope of CO_2, NH_3, H_3PO_4 and H_2SO_4. Each plot shows in how many cases a single deletion reduces the scope size by a given number of compounds. The diagram reveals that in all four cases the majority of such deletions does not affect the scope size at all. Most of the other deletions have only a small effect on the scope size. However, there are a few reactions whose deletion significantly reduces the scope sizes.

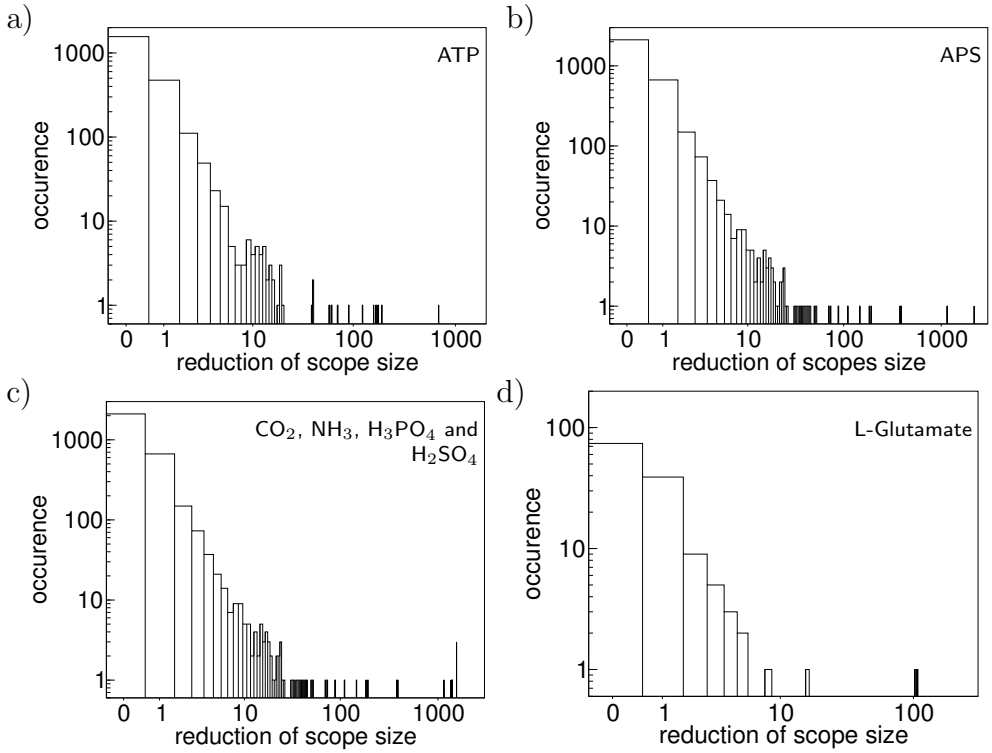


Figure 4.1: The effect of single reaction deletions on the scope size. It is shown, how many single reaction deletions reduce the scope size by a certain number of compounds. a) for the scope of ATP, b) for the scope of APS, c) for the scope of CO_2, NH_3, H_3PO_4 and H_2SO_4 and d) for the scope of L-glutamate.

Specifically, the network resulting from the expansion of ATP contains 1554 compounds and 2328 reactions. As shown in figure 4.1a, for 1589 reactions, a single deletion does not influence the scope size. For another 718 deletions, the reduction of the scope size is smaller than 20. Among the 21 reactions whose deletions have a larger effect, there exist two which are

very critical for the scope size. The deletions of these reactions result in a reduction of the scope size by 689 and 690 compounds, respectively. A closer inspection reveals that these two reactions are

$$\text{L-Aspartate} + O_2 \rightleftharpoons \text{Iminoaspartate} + H_2O_2, \tag{4.5}$$

catalyzed by the enzyme L-aspartate oxidase (EC 1.4.3.16) and

$$\begin{aligned}\text{Pyridine-2,3-dicarboxylate} + 2H_2O + \text{Phosphate} \rightleftharpoons \\ \text{Iminoaspartate} + \text{Glycerone phosphate},\end{aligned} \tag{4.6}$$

catalyzed by a carbon lyase (EC 4.1.99.-). Deletion of one of the two reactions disables the only non-NAD-dependent synthesis path of NAD from ATP in the network. Hence, NAD and many other compounds cannot be synthesized from ATP anymore.

The fact that there exists a small number of reactions whose deletion affects the scope size dramatically, is also visible in the other examples shown in figure 4.1. In the case of APS, a deletion of the reaction

$$\text{APS} + H_2O \rightleftharpoons \text{AMP} + \text{Sulfate}, \tag{4.7}$$

which is catalyzed by the enzyme adenylylsulfate sulfohydrolase, results in a total collapse of the scope (figure. 4.1b). This devastating effect can be explained by the fact that this reaction is the only one within the network which can metabolize the compound APS using no other compound except water. The elimination of this reaction stops the expansion process at the very beginning.

As mentioned above, the scope of CO_2, NH_3, H_3PO_4, H_2SO_4 is exactly the same as the scope of APS. As expected, the analysis of the effect of single deletions yields similar results in both cases (cf. figure 4.1b and c). However, there exist differences for those reactions which have a strong impact on the scope size. The analysis of the robustness of the scope of L-glutamate, which is significantly smaller than in the other examples (428 compounds associated with 514 reactions), results in a similar behavior (figure 4.1d).

4.3 Robustness against multiple deletions

As a next step, the effect of the simultaneous removal of more than one reaction is examined. Figure 4.2 shows how the size of the ATP scope decreases with an increasing number of removed reactions. Specifically, a two-dimensional distribution is shown with the shading indicating the probability

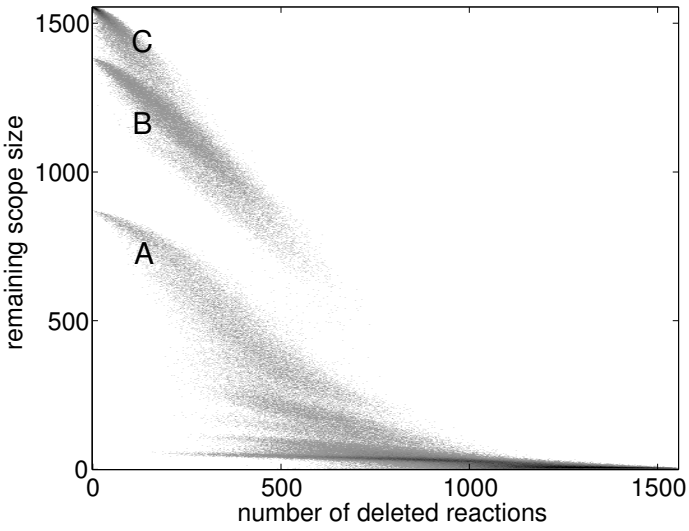


Figure 4.2: Effect of multiple reaction deletions on the size of the ATP scope. The shading gives the probability that the deletion of a certain number of reactions (x-axis) leads to a reduced ATP scope of a certain size (y-axis). For each x-value, the ATP scope has been calculated on 1000 different networks, randomly reduced by the corresponding number of reactions

that a reduction of a given number of reactions results in a certain scope size.

There exist distinct domains in which these probabilities are high. These result from the deletion of reactions already identified as critical in the analysis of single deletions, as shown in figure 4.1a. Domain A contains cases in which one of the two reactions (4.5) and (4.6) has been deleted, resulting in a reduction of the scope by at least 689 compounds.

For domain B a group of reactions is responsible which is shown in figure 4.1a at a reduction of around 190 compounds. Hence, scopes in domain B are reduced by about 190 or more compounds.

Domain C on the other hand only contains cases where no such critical reactions are removed. In this case, the scope size almost linearly decreases with a decreasing network size. Clearly, with an increasing number of deleted reactions the probability that a critical reaction is deleted increases. Therefore, the domains B and C get depleted in favor of domain A for large numbers of deleted reactions.

Similar behavior can be observed for other seed compounds, too. It can be concluded that scope sizes are critically influenced only by a few reactions. The removal of other reactions reduces the scopes sizes only in the

same manner as the network size is reduced. Also, most critical reactions apparently show a critical effect only for certain seeds. Hence, the scopes are generally robust against the deletion of reactions.

Moreover, the behavior of multi scopes can be analyzed in dependence of the underlying network. In figure 4.3 scope size distributions for networks randomly reduced by a certain number of reactions are shown. The shading indicates the propability that a scope of a random seed has a particular scope size when calculated on a network from which a certain number of reactions has been deleted. For zero deleted reactions the distribution is actually similar to the distribution shown in figure 3.9. Here, for each number of deleted reactions, 25 random networks were generated and for each of these networks 400 scopes of 30 random seed compounds were calculated. The bands of multi scopes which can be attributed to characteristic scopes like the scopes of APS or ATP, persist over a wide range of network sizes, even though they start to blur after about 300 deleted reactions. The scope sizes in these bands descend almost linearly with the number of reactions deleted. Apparently, also here critical reactions exist leading to the emergence of domains between the bands of the characteristic scopes. At some point (around 50 percent of all reactions deleted) the network seems to disintegrate, not allowing for the existence of considerable scopes.

4.4 Effects on the scope hierarchy

The scope size distribution suggests that characteristic scopes persist over a certain range of network modifications and that their sizes scale approximately linearly with the size of the network. In this section, the structure of the scope hierarchy itself is analyzed in dependence of a changing underlying network.

In chapter 3 it has been shown that an artificial network containing all possible reactions results in a simple and intuitive hierarchy where all scopes represent a certain set of building blocks. When a certain number of reactions is randomly removed, the scope hierarchy gets transformed into a new structure, where characteristic scopes, represented by nodes with large degree, still characterize a specific set of building blocks, while others are already too loosely connected to make use of their building block content and are represented by nodes with a small degree or by sink nodes. In the following, the transition between these two hierarchies is evaluated by observing the scope hierarchies of a network when consecutively more and more reactions are deleted. This process is then continued to a case where all reactions are removed.

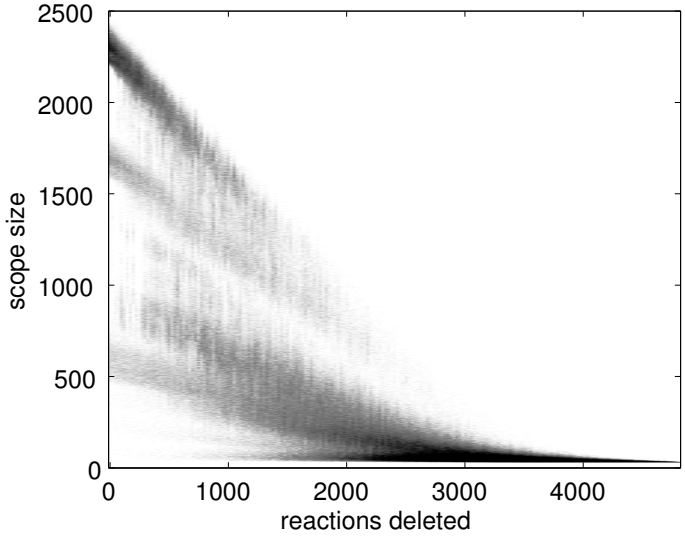


Figure 4.3: Effect of multiple reaction deletions on the sizes of random multi scopes. The shading scales with the number of scopes having a certain scope size and were calculated on a network reduced by a certain number of reactions. Here, for each number of deleted reactions (step size 25), 25 random networks were generated and for each of these networks 400 scopes of 30 random seed compounds were calculated. Therefore, the minimal scope size is 30.

Figure 4.4 shows 10 selected steps of such a process for an artificial network defined by $N_{(A,B,C,D,E)}$=(4,3,2,1,1) which contains 239 compounds and in its unreduced form 3816 reactions. Figure 4.5 gives characteristic values of the graphs. The unreduced network yields the expected simple hierarchy. Many seed compounds result in the same scopes, i.e. are interconvertible. Interestingly, a large number of reactions can be removed before this hierarchy is substantially changed. In the next phase, the number of nodes and ranks in the graph increases. The scopes of the original simple hierarchy persist and become the so called characteristic scopes, while new, less connected scopes appear, resulting from seed compounds which are, due to reaction deletion, not anymore able to reach one of the characteristic scopes. The characteristic scopes persist over a large number of deletion steps.

At some point, the number of ranks in the hierarchies becomes again smaller while the number of nodes approaches the total number of compounds, which means that there are almost no interconvertible compounds anymore. Also, the number isolated nodes strongly increases. These effects are due to a beginning disintegration of the network whose ability to do con-

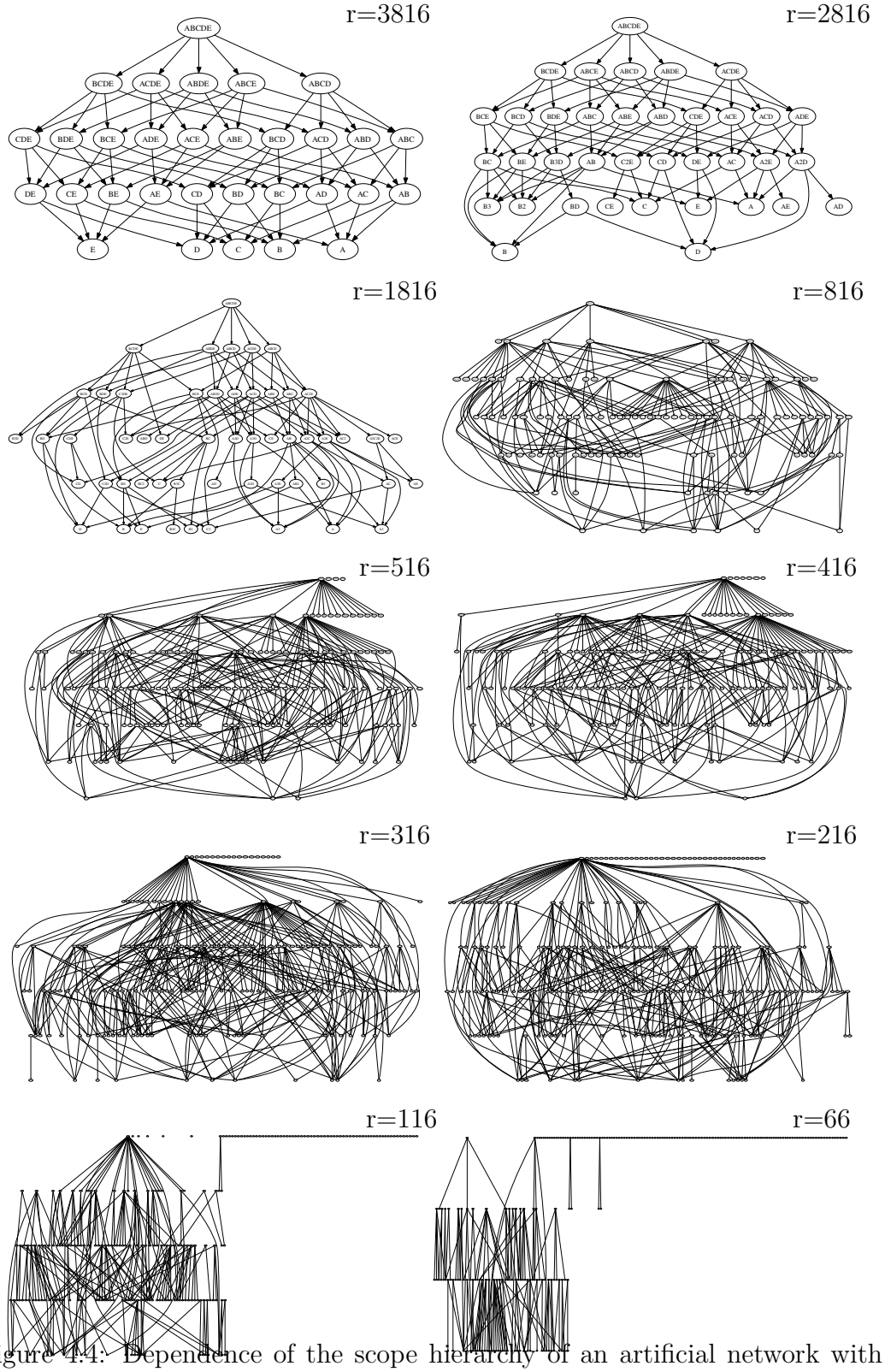


Figure 4.4: Dependence of the scope hierarchy of an artificial network with $N_{(A,B,C,D,E)}$=(4,3,2,1,1) (239 compounds, maximal 3816 reactions) on the number r of remaining reactions in the network.

versions is strongly reduced. This eventually also leads to a disappearance of the characteristic scopes which is consistent with the effect observed for the scope size distributions in figure 4.3. Eventually, the network disintegrates completely when all reactions are removed.

The source-sink connectivity (figure 4.5) gives an idea of the ability of the network to do conversions. In a complete network there exists a scope containing all building blocks and therefore all other scopes are included in this top scope. For the hierarchy graph that means that there exist one source and as many sinks as there exist building blocks and the source is connected to all sinks. The source-sink connectivity is defined as the quotient of the number of all connected source-sink-pairs and the number of all in principle possible source-sink-pairs, or more precisely

$$c_{sd} = \frac{p + n_i}{(n_s + n_i)(n_d + n_i)}, \tag{4.8}$$

where p is the number of connected source-sink-pairs (not considering isolated nodes), n_s the number of sources, n_d the number of sinks and n_i the number of isolated nodes, which can be interpreted as source and sinks at the same time and therefore influence the source-sink connectivity in the above described way. For the hierarchy of the complete network this connectivity is 1. In the completely disintegrated network all compounds can only be converted into themselves, which means that the hierarchy contains only isolated nodes. Therefore the source-sink connectivity is n/n^2, with n being the number of compounds in the network, i.e. for the analyzed network 0.0042. For all cases between these two extrema the source-sink connectivity monotonously decreases with decreasing numbers of reactions in the network.

A similar procedure can analyze the effect of network modifications on the scope hierarchy of the KEGG network. However, it should be noted, that the addition of new artificial reactions to the KEGG network is methodically difficult. Therefore, only a further reduction from the present network is analyzed. It is apparent, that the KEGG network is not a complete network in the sense that it is able to perform all possible conversions. It can therefore be expected that the process on the KEGG network only covers part of the process shown on the artificial network.

Figure 4.6 shows the graph characteristics of the hierarchies of the KEGG network during the reduction process. When comparing with the characteristics of the artificial network (figure 4.5) it becomes clear, that the KEGG network starts somewhere in the middle of the process observed on the artificial network. Afterwards, the reduction processes proceed similar.

Apparently, the original KEGG network has already been reduced to a point where further reaction deletions have a significant influence on the

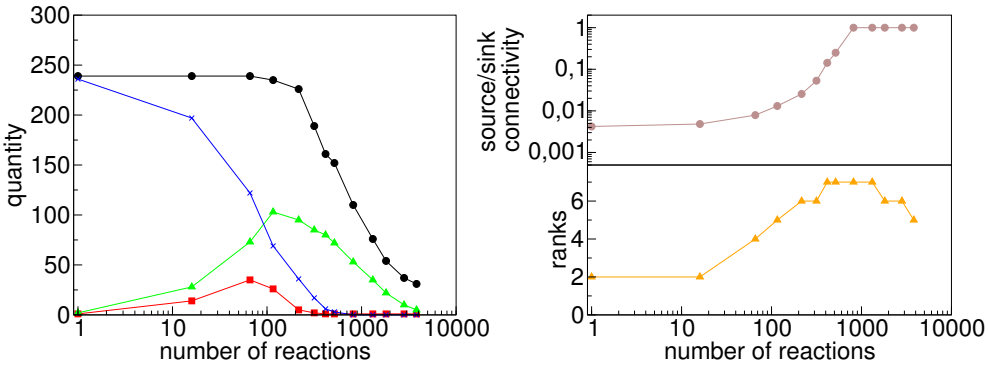


Figure 4.5: Characteristics of the scope hierarchies in dependence of the number of remaining reactions in the network with $N_{(A,B,C,D,E)}$=(4,3,2,1,1). Left graph: the total number of nodes (circle), the number of source nodes (squares), the number of sink nodes (triangles) and isolated nodes (x). Right graph: number of ranks in the hierarchy (triangles) and the source-sink connectivity (circles), i.e. number of connected source/sink pairs divided by the product of source and sink vertices. It should be noted that the reduction process actually proceeds from right to left since the number of reactions in the network is decreased.

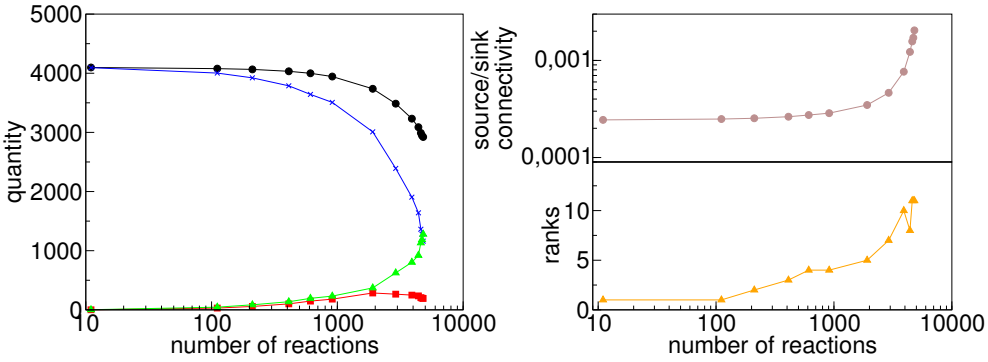


Figure 4.6: Characteristics of the scope hierarchies in dependence of the number of remaining reactions in the KEGG network. Left graph: the total number of nodes (circle), the number of source nodes (squares), the number of sink nodes (triangles) and isolated nodes (x). Right graph: number of ranks in the hierarchy (triangles) and the source/sink connectivity (circles), i.e. number of connected source-sink pairs divided by the product of source and sink vertices.

structure of the hierarchy, but where the ability of performing conversions is still large enough to show characteristic scopes. This seems to be a reasonable compromise between a network having too many redundant reactions and a network which only has a limited ability to perform conversions.

4.5 Irreversible reactions

So far calculations have been made assuming all reactions as reversible. This section will give some argumentation why this is useful for the analyses presented in this work and will also show how the results are affected if information about irreversibility is included.

In principle, all reactions are reversible. The question whether a reaction occurs in forward or backward reaction is determined by the sign of the Gibbs-Free-Energy change of that reaction

$$\Delta_r G = \Delta_r G^0 + RT \ln \prod c_i^{\nu_i}. \tag{4.9}$$

The c_i are the concentrations of the metabolites and the ν_i their stoichiometric coefficients which are negative for substrates. $\Delta_r G$ is the work that can be obtained from the reacting system at constant temperature and pressure when transforming the substrates into products. If the concentrations are adjusted in a way that $\Delta_r G$ is zero, no work can be obtained and hence no transformation can proceed spontaneously. The system is thus in equilibrium.

$\Delta_r G^0$ is the difference of the standard Gibbs free energy of formation $\Delta_f G$ of the products and the substrates

$$\Delta_r G^0 = \sum_j \nu_j \Delta_f G_j, \tag{4.10}$$

with ν_j being the stoichiometric coefficient of the jth metabolite. The $\Delta_f G$ can be determined experimentally and depend on the changes in enthalpy and entropy in a system when the corresponding metabolites are formed from their elements at constant temperature and pressure at standard conditions

$$\Delta_f G = \Delta_f H - T\Delta_f S. \tag{4.11}$$

While the $\Delta_r G^0$ of any reaction is fixed, different concentrations of the participating metabolites may result in different signs of $\Delta_r G$. Therefore the direction is not predefined. More details on this can be found in Atkins [1990].

For specific organisms, the concentration of metabolites may be limited to a certain physiological range. Hence, situations may occur in which a reaction

can only proceed in one direction and it may be useful to incorporate this directionality in the scope calculation.

Clearly, the directionality is not a structural information anymore. For its determination thermodynamic properties like the standard Gibbs free energy along with physiological ranges of concentrations have to be known. This requires a much more detailed knowledge on the metabolic networks which is yet not available on a large scale.

The calculations done so far therefore utilized reversible reaction only. This is not problematic as all considerations have been done on a general, non organism specific level. Hence, constraints to metabolite concentrations cannot easily be made. Further, especially for the scope hierarchies, structural properties of the compounds are analyzed which are independent of any concentrations or reaction directions.

However, for manually curated networks, where the necessary information is available it is certainly useful to incorporate directionality into the scope calculations. In this chapter scopes are calculated on the reference network using the directionality information as provided by KEGG. Even though this information is certainly not the most complete, it is sufficient to demonstrate the methodology.

Figure 4.7 shows the distribution of scopes sizes for single and multi scopes. General structures observed in the reversible network, like bands of multi scopes next to characteristic single scopes, also exist in the network with directionality incorporated. As a tendency, scopes sizes are smaller than in the original network, as the introduction of irreversibility generally decreases the ability of the network to do conversions.

Also, interconvertibility is lost. While most characteristic scopes can still be produced from some seed compounds, other, formerly interconvertible compounds yield now smaller scopes. For example the characteristic scope of ATP still exists and can be produced from a variety of compounds, like ATP,ADP, NAD, etc, but some compounds, like GDP and UDP now fail to produce the same scope.

The largest characteristic scope in the reversible network is split up in two distinct scopes in the case of irreversibility. While APS and PAPS still reach a scope size of 1902, the scope of Dephospho-CoA only contains 1642 compounds. The reason is that with irreversible reactions, sulfur cannot be extracted from CoA. This makes CoA a separate building block, if no other sulfur is available. The frequent occurrence of CoA in other compounds and the number of compounds whose production require this cofactor make its scope considerably larger than the next smaller subscope, the scope of ATP.

The largest characteristic scope is again the scope representing the chemical elements C, N, P and S. Additionally to the compounds of the scope of

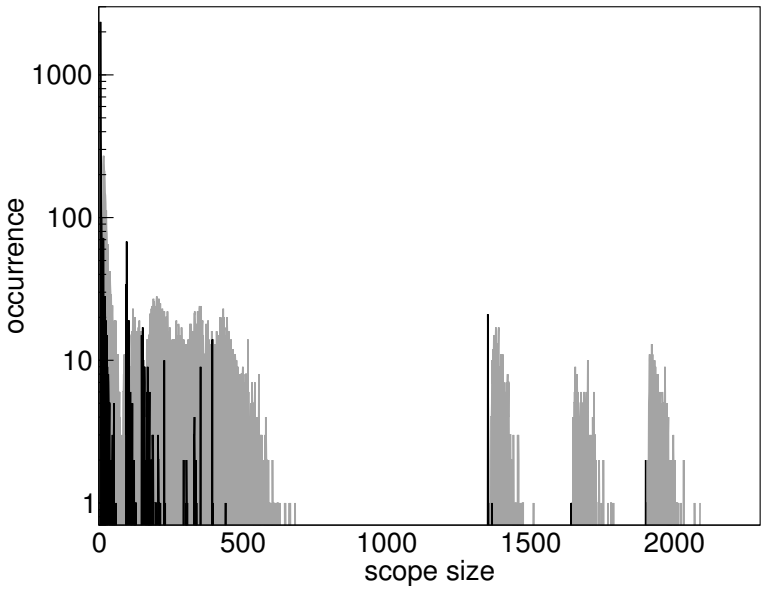


Figure 4.7: Distribution of scope sizes of 10000 scopes of 15 random seed compounds calculated on the KEGG network with irreversible reactions (gray curve). The distribution of single scopes is given in black.

Dephospho-CoA, this scope holds sulfur containing compounds including the cofactor S-Adenosyl-L-methionine which is involved in methyl group transfers. This scope is again substantially larger than the scope of Dephospho-CoA.

4.6 Analysis of organism specific networks

Different metabolic networks also occur if the networks of different organisms are analyzed. These networks have been adapted by evolution to the ecological niches in which the corresponding organisms live in. On the other hand, due to their ancestry, the networks of different organisms show many similarities.

As stated in this work, the scopes define functional measures of the metabolic capabilities of the analyzed metabolic networks. The differences in the network structure among different organisms may or may not result in different metabolic functionalities. In particular, it may be the case that alternative synthesis routes in different networks synthesize the same metabolic products, leading to the same or at least similar scopes. However, in general it can be expected that different organisms show different metabolic functionalities and hence different scopes.

As an example, the scope of ATP is analyzed for all organism specific networks defined in the KEGG database. Figure 4.8 shows the distribution of scopes sizes in dependence of the number of reactions in the corresponding organism network. The membership to one of the three domains of life is indicated.

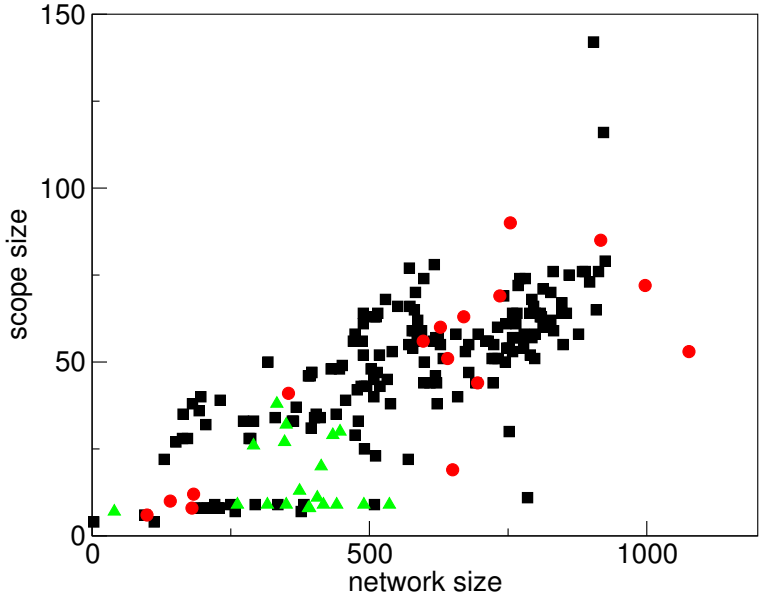


Figure 4.8: The size of the scope of ATP for different organisms is plotted against the number of reactions in the corresponding network. The membership of a species to one of the three domains of life, bacteria (squares), eukaryota (circles) and archea (triangles) is indicated.

Clearly, the scope of ATP is different for different organisms. This differences cannot only be explained by the different sizes of the specific networks. For this particular example, there apparently exist two groups of organisms, where the members of one group can use ATP to produce a significant number of compounds, whereas members of the other group cannot.

The concept of scopes can hence be utilized to perform an comparative analysis of the metabolic capabilities of different organisms. Certainly, instead of ATP, more realistic resources should be used. For this example, ATP has been used as it covers substantial parts of the metabolic networks which ensures that the resulting functional measures reflect properties of the whole network and not only of a small part.

A thourough analysis of organisms specific networks using the concept of scopes goes beyond the scope of this work. This path is further followed in Ebenhöh et al. [2005] and Ebenhöh et al. [2006].

Chapter 5

Discussion

In this work, the concept of scopes has been applied to large scale metabolic networks. The concept as such predicts potential products which can be synthesized from seed compounds. Hence, the scopes are functional measures, describing the synthesizing capacity of the underlying metabolic networks, given the availability of predefined external resources.

The concept is based on an intuitive algorithm which can be implemented in a fast and efficient way. Therefore it can be used for a systematical analysis of the functions of metabolic networks, exploring a vast number of resource constellations or network variants like mutant networks or different organisms.

The method is purely structural, allowing for its application on larger scale networks for which kinetic parameters are often unknown or at least incomplete.

5.1 Summary of results

First, the scopes of all 4104 single seed compounds have been calculated on the organism independent KEGG network. Interestingly, these single scopes can already become quite large. The scope of APS, the largest scope, covers approximately 50% of the network.

Further, the distribution of scope sizes is very inhomogeneous, showing larger peaks and long gaps especially for larger scopes. The large peaks can be accounted to the existence of groups of interconvertible compounds which all produce the same scope. Scopes seem to cluster next to prominent peaks which leads to large empty areas between them.

The concept of interconvertibility also applies to scopes of more than one seed compound. These multi scopes may coincide with single scopes as shown

for the example of CO_2, NH_3, H_3PO_4 and H_2SO_4 which yields the same scope as APS.

The expansion process itself can reveal valuable information about the topology of the metabolic network. When calculating large scopes, large parts of the network are crossed during the expansion. In that way local dependencies of the participating compounds and reactions can be analyzed. It has been shown that certain compounds trigger an acceleration of the reaction attachment in subsequent steps. In particular, the cofactors ATP/ADP, NAD^+/NADH, $NADP^+$/NADPH and coenzyme A have been identified to produce the described effect which can be accounted to the high number of reactions those cofactors take part in. However, the effect can only be observed if at the point of the cofactor incorporation a substantial number of substrates has already been included in the network.

It is further possible to define the vicinity of compounds by the number of steps the expansion process needs to proceed from the seed to a certain metabolite. Clearly, the distance between a seed compound and a target metabolite is dependent on the other compounds in the seed. However, these other seed compounds can be chosen as needed in a particular context. This definition of distance is more realistic than graph theoretical distances (cf. section 1.2) as it avoids shortcuts via highly connected compounds which do not reflect a real transport of chemical content. This goes in line with the findings in Arita [2004] where the flow of carbon atoms through metabolic networks is investigated.

Furthermore, the behavior of scopes has been analyzed assuming that cofactor mediated reactions can occur even if these cofactors cannot be produced by the seed. Generally, in that case scopes become larger and more compounds can be interconverted into one another. In particular, scopes of compounds containing the elements C,H and O have been strongly increased in size.

In addition to the scope, which is the set of metabolites which can be synthesized from the seed, it can also be investigated which seeds can produce a desired set of target metabolites. First, it was analyzed which seeds can reproduce the complete network. It turned out that there exist many small parts in the surroundings of the network requiring their own specific seed compounds. As a result, more than 500 seed compounds where needed even though large parts of the network can be covered using only a few of those. The latter parts have been called the central regions of the network.

The synthesis of these central parts has been analyzed in detail. A target set has been defined, containing compounds which can be found in the majority of all organisms. It can be assumed that this set contains many essential compounds necessary for cell maintenance and growth. To produce

this target set, a large number of combinations of seed compounds is possible albeit the total number of compounds in a particular seed was only 4 in average.

The scopes of single seed compounds can be represented as a hierarchy. Super scopes and their included sub scopes are represented as predecessors and successors in this hierarchy. Consequently, larger scopes tend to reside on top of the hierarchy while smaller scopes are situated rather at the bottom.

The position in the hierarchy appears to be connected to the chemical content of the scopes and hence their seed compounds. For example, the scopes of Adenine, Adenosine, ATP and APS are subsequently included into one another, which reflects the fact that each of compounds in the list contains a distinct chemical group less than the next compound.

In the hierarchy graph, where each node represents a distinct scope, certain nodes show a large in or out degree. The corresponding scopes can often be reached by a large number of interconvertible single seed compounds. The analysis of these seed compounds showed that they contain a certain set of chemical elements or chemical groups in common. In particular, these scopes contain the largest fraction of compounds of a certain composition and were therefore called characteristic scopes.

Scopes of multiple seed compounds also show an inhomogeneous size distribution. For large sizes they tend to be a bit larger than certain characteristic single scopes leading to bands to the right of these scopes in the size distribution. It has been shown that multi scopes in these bands actually contain the corresponding single scope as a sub scope. Hence, the scopes contain the chemical content of the corresponding characteristic scopes as well as some other less potent content which explains their slightly larger sizes.

In order to test the assumed connection between the chemical content and the structure of the scope hierarchy, a simple model of an artificial chemistry has been introduced which basically ensures the conservation of artificial building blocks in its reactions.

For a network containing all possible conversions, each scope represents a specific set of these building blocks. The corresponding hierarchy has a node for each building block combination. Nodes of the same rank have the same number of building blocks and successors have exactly one building block less than their predecessors.

When a large number of reactions is randomly removed, the corresponding scope hierarchy of single scopes transforms into a structure similar to the scope hierarchy observed for the KEGG network. Also here, characteristic scopes can be identified by their high degree and assigned to a specific building block combination. These are the remains of the clear hierarchy of the

complete artificial network. Additionally, there exist now a large number of less connected scopes which reflect the declining ability of the network to do conversions.

As in the KEGG network, multi scopes in the artificial network may coincide with single scopes with the same chemical composition. This is always the case for the complete network and to a less extent also for the randomly reduced version. If the building blocks are used in an uneven manner, multi scopes tend to cluster next to characteristic scopes in the same way as observed for the KEGG network.

It has been analyzed in detail, how the ability of the network to perform conversions is affected by the consecutive removal of reactions. Starting with the complete artificial network, a large number of reactions can be removed before the hierarchy starts to change. Then the number of nodes in the hierarchy increases and a subhierarchy of high degree characteristic scopes emerges. Further reduction leads to a disappearance of this subhierarchy, indicating a disintegration of the underlying network. Altogether, this transition phase is rather short and happens rather in the end of the reduction process. It has been shown that the KEGG network is situated in this transition phase. A reduction of this network also eventually leads to a disappearance of the hierarchy.

Furthermore, the dependence of the synthesizing capacity on changes in the topology of the KEGG network has been studied. Clearly, the removal of reactions generally leads to a reduction of the scopes, but may also have no effect, if the missing reactions can be compensated by alternative routes.

The analysis of the deletion of single reactions showed that the KEGG metabolic network has in most cases the capability to compensate the absence of single reactions. There exist a few reactions which may influence the scope size dramatically. However, these reactions cannot be considered as generally critical to the whole network as their dramatic effect is often specific to particular seeds.

The analysis has been extended to random multiple deletions on the example of a relatively large scope, the scope of ATP. It has been shown that the effect is small as long as the formerly identified reactions with the large single effect are not chosen. Of course, the scope size does not remain the same if a large number of reactions is removed. The scope size is in general reduced in the same way as the corresponding network size. This behavior can be interpreted as robust. The analysis has further been extended to random multi seed scopes. Here the whole structure of the scope distribution more or less uniformly scales down with the network size.

5.2 The synthesizing capacity

The straightforward application of the concept of scopes allows to answer the question which compounds are possible metabolic products of a particular network provided with a given set of nutrient compounds.

Scopes can predict whether a particular organism can produce desired target metabolites from given resources. This may be interesting for biotechnological applications where particular target metabolites are to be produced by cultured organisms. It should however be mentioned that such applications usually require a production at steady state, a high yield and a high throughput, which the scopes alone cannot guarantee.

The scopes describe the metabolic processes during cell growth particularly well. In such a case a constant replenishment of all metabolites is required and thus a steady state flux through all reactions in the fully expanded network can exists. Hence, it is possible to check whether an organism is able to produce the metabolites required for maintenance and growth from the provided resources.

The results can be refined by including more biological information into the calculations, if such data is available. In particular, information about compartmentalization and transport processes, as described in the introduction, will increase the accuracy of the results. It is also possible to incorporate the current regulatory state of the metabolism if this data can be reduced to define the reactions as active or inactive. This has been done recently in Ebenhöh and Liebermeister [2006]. Clearly, this can only be done statically. In particular, feedback of the metabolism to the regulatory pathways cannot be modelled by this method.

The method of network expansion generally assumes that all metabolites but the seeds have initially a zero concentration. This is necessary as otherwise these metabolites would also be used as seeds for the production of the scope. However, real cells are not empty. Usually such metabolites present in cells can only be used for a continuous synthesis of other products if they are themselves producible from the nutrients. There exist however cases where this replenishment is done externally to the pathway currently under investigation. This may in particular be the case for cofactors. Generally these are held on a constant level by the cell. Also, these do not contribute to a large extent to the product mass but rather mediate the participating reactions.

If such cofactors are missing, the scopes may predict less products than the cell could actually produce. Hence, a special treatment of these is necessary. The expansion process is modified in order to reflect the presence of specific cofactors as catalysts of reactions. It is ensured that the cofactors itself are not used as substrates for the synthesis of other metabolites. As

described, scopes may become significantly larger if cofactor functionalities are considered. Hence, even though chemical reactions can in principle also occur without cofactors, for example the direct uptake of phosphate instead of an ATP catalyzed reaction, many important reactions apparently exclusively depend on the presence cofactors. This confirms the importance of cofactors in metabolic networks.

The concept of scopes can also be utilized to determine seed compounds from which the analyzed networks can synthesize certain target compounds. Also here the above mentioned refinements will improve the accuracy. The analysis is most useful if performed on organism specific networks. If the target set is chosen to contain metabolites necessary for cell maintenance and growth, the results can be used to determine potential growth media for the analyzed organisms. The target set and the set of preferred seed compounds which defines the compounds that can be taken up by the organism have to be carefully chosen in order to obtain the most exact results. This methodology has been recently applied in Handorf et al. [2007].

As discussed in the introduction, due to the purely structural nature of the scopes, predicted products are putative and kinetic constraints may further confine the set of possible products. Furthermore, metabolic processes are strongly regulated, modulating the activity of the involved enzymes and thereby controlling the diversity of products.

It is therefore an intriguing question which biochemical information can be drawn from the scopes themselves. It is especially useful to apply the methods described in this work whenever a manifold of networks or resource combinations is to be analyzed. The calculations can provide quick estimates of metabolic capabilities. Such investigations may include the search for organisms capable of producing certain target compounds in a biotechnological process or the determination of possible growth media on which the organisms in question can live. The algorithm will reduce the list of candidates significantly. It can however generally be expected that more candidates, i.e. products or seed combinations, are returned than a more detailed method would allow. Hence, a more precise but also more time consuming method can subsequently be applied, testing which of the candidates actually fulfill the corresponding requirements.

Scopes depend on the structure of the analyzed metabolic network. The structure of networks is determined biologically, i.e. by the organsism, the mutant or the pathway under investigation and technically by the provided data and hence indirectly by the method of data generation like sequence alignment and gene annotation. For a biological analysis the biological effect is desired, providing the possibility to analyze the different metabolic capabilities of different organisms or mutants. However, the technical effect is not

desired, as for such investigations the analyzed network is ideally the actual network of the analyzed organism.

Therefore, in this work, the effect of changes in the network structure on the scopes has been studied. As a result it turned out, that the scopes calculated on the KEGG network are generally robust against the random removal of reactions. This is even true if several reactions are removed consecutively. In this case, the scope sizes generally scale with the size of the remaining network.

There exist only a few reactions, whose removals affect the scope size dramatically. This effect is however dependent on the specific seed and does not affect the general robustness of the network to a high degree.

This result shows that potential errors in the networks imported from the KEGG database will in general not affect the results presented in this work dramatically. Hence, if the calculation will be repeated with later improved data, the conclusions drawn in this work will mostly remain the same. A comparison of scope size distributions of two KEGG versions as shown in the appendix A.2 confirms this statement.

Moreover, the scopes can also help to improve the metabolic data. If for an organism information on nutrients and metabolic products is available, the failure of the scope to predict a certain product may indicate missing reactions in the network. Clues on the location of such missing reactions can be obtained by determining the seeds which would enable the organism to produce the initially non-producible product.

As mentioned, the investigation of network modifications is also of biological interest. The general robustness of the KEGG network reveals that there exist many alternative routes which can fill in if certain reactions fail. Clearly, as this analysis is done on the organism independent reference network of the KEGG database, the robustness of organism specific networks still has to be determined. Ebenhöh et al. [2005] provides a closer examination of this topic.

Also, by removing reactions from the network, a large number of potential mutants can be generated. With the described methods it is possible to analyze the synthesizing capacities of these mutants and hence predict their viability or identify critical reactions in their networks.

5.3 Building blocks

Apart from metabolic predictions, as discussed in the last section, the concept of scopes can be used to analyze the structure of metabolic networks and to formulate hypotheses about principles which determined this structure

during evolution.

Metabolic networks are obviously shaped by the chemical structure of their metabolites. Two metabolites can only be interconvertible if they possess the same chemical content. This content is defined by the set of chemical groups or chemical elements in the metabolites. Clearly, two such compounds must have the same chemical elements, as these can neither be created nor annihilated by chemical reactions. The two compounds may however also share common chemical groups. In fact, they must share a specific group if this group is conserved in the network, meaning that there is no reaction which assembles or disassembles the group.

The chemical content imposes a hierarchy on the scopes. In this hierarchy, scopes containing more chemical content are superordinated to scopes possessing less content. Seeds and their scopes contain the same chemical content (cf. equation 1.21 indicating that a scope is interconvertible with its seed) and hence also a hierarchy on the seed metabolites is inferred. The resulting hierarchy graph can be interpreted as an alternative view on metabolism, specifically highlighting the chemical richness of the participating compounds.

The coherence of the hierarchy and the chemical content has been confirmed in artificial metabolic networks. Here, compounds hold building blocks, which may represent chemical elements or conserved chemical groups. In particular, this analysis revealed the connection between the building blocks and specific prominent scopes, the so called characteristic scopes. Clearly, if the network contains all in principle possible reactions, all compounds with the same building blocks must be interconvertible and will be represented by a single node in the hierarchy graph. If not all conversions are included, interconvertibilities are lost. However, over a wide range of reaction deletions, the majority of compounds with the same building blocks remain interconvertible. The scopes of these compounds become the characteristic scopes. Those compounds not anymore interconvertible with other compounds yield new and distinct scopes. The corresponding nodes in the hierarchy are often directly connected to one or more of the characteristic scopes which in turn explains the high degree of these.

The characteristic scopes can also be identified in the KEGG network. High degree nodes in this hierarchy correspond to scopes containing a specific set of chemical elements. These also possess a large number of interconvertible seed compounds. In particular, the characteristic scopes representing combinations of the elements C,N,P and S, except CS could be identified. Even though there exist compounds with the element combination CS, these apparently do not form larger groups of interconvertible compounds. Also, scopes containing only N or P do not possess a large number of seed com-

pounds. These two peculiarities may be accounted to the fact that not all combinations of elements have the same probability to constitute biologically relevant metabolites. Absolute frequencies of compounds with specific chemical elements can be found in appendix A.5. It should again be noted that the conservation of the elements H and O cannot be seen in the hierarchy, as water has always been included in the seed.

In the KEGG hierarchy there exist also characteristic scopes which can be accounted to the existence of specific chemical groups. All interconvertible seed compounds producing such a scope contain the corresponding group. Two examples have been analyzed, the scopes of Arachidonate and Retinal. The synthesis of Arachidonate is not included in the data set and hence Arachidonate and its products, the leukotrienes and prostaglandins contain at least one own conserved building block.

The situation for Retinal is somewhat different. It contains the chemical element C (along with H and O) only (chem. formula: $C_{20}H_{28}O$). It is not interconvertible with most other compounds containing only C (and possibly H and O). However, there does not exist a specific conserved chemical group in Retinal. Retinal can actually be produced from ATP which does not contain any group similar to Retinal. There exists however a conserved group if only reactions are considered which exclusively use compounds containing the elements C,H and O. As described, for the synthesis phosphorylated intermediates (e.g. Isopentenyl-PP) are necessary which additionally contain the element P. This leads to the existence of a separate characteristic scope for Retinal.

Such a partial conservation of chemical groups generally implies that the production of the group, and if applicable also their degradation, proceeds via chemically more complex intermediates. Equally, the need for cofactors may define a partially conserved building block. Clearly, if a cofactor is not present, certain compounds may not be produceable from compounds with the same chemical content. However, other, chemically more potent compounds may able to produce the cofactor as well as the desired compound, indicating that no additional strictly conserved building block is involved.

This behavior effectively defines subnetworks in which certain chemical groups are conserved. Such a subnetwork is surrounded by reactions which utilize compounds with a larger chemical content or are dependent on cofactors.

As argued before, many cofactors are ubiquitous in the cell. Hence, the cofactor-dependent partially conserved building blocks may not play an important biological role as the necessary reactions can generally operate. To address this, the analysis is also performed assuming that the functionalities of certain cofactors is present. In fact, the corresponding hierarchy unifies

many different nodes indicating the increased capacity of the network to do conversions.

On the other hand, the existence of partially conserved building blocks due to more complex intermediates may have a biological meaning. In particular the ligation of phosphates to intermediates may indicate special energetic or regulatory needs in the synthesis of a specialized chemical group. In that way, the occurrence of a partially conserved building block in the hierarchy may indicate a special role of the metabolites containing it.

Generally, the conservation of building blocks depends on the reactions in the network. While the conservation of chemical elements in the hierarchy graph is due to the principal inability of chemical reactions to convert elements into one another, the conservation of chemical groups depends on the ability of the network to synthesize or degrade such groups. If such reactions are missing for a specific group, this group becomes a conserved building block.

In a broad interpretation, each node in the hierarchy graph represents a unique combination of strictly or partially conserved building blocks. The special role of characteristic scopes would then be that they represent a popular building block combination which many compounds share. The other scopes contain building blocks which occur in only a few compounds. For such less frequent building blocks it cannot be assumed that they occur in various combinations with other building blocks. Consequently, they will not be part of a clear subhierarchy of characteristic scopes as seen specifically for the chemical elements C,N,P and S which occur in all combinations in the network. It can rather be expected that the manifold of implicitly defined building blocks together creates the background of non characteristic scopes observed in the hierarchy graph.

The chemical content is also visible in the results of the seed prediction. This analysis has been done for the seeds of central parts of the metabolism and for the seeds of the complete network.

Only a few seed compounds were needed to produce the central metabolites. Still the scopes of these seed compounds cover more than 50% of the whole network which is comprehensible if considering that a single, but complex compound like APS alone covers a similar fraction of the network. From the chemical structure of the calculated seed compounds it can be seen that each compound actually provides one or more chemical elements, rather than specific chemical groups. It should be noted that the elements themselves are not among the seeds, as they are usually metabolically difficult to access. These findings indicate that at least the central region of the KEGG network is autotroph, meaning that all compounds can be synthesized from small inorganic compounds, like CO_2, NH_3, H_3PO_4 or H_2SO_4.

To produce all compounds of the KEGG network, in average 534 seed compounds were needed. As each seed compound must provide at least one additional conserved building block, 534 gives a lower limit for the number of conserved building blocks in the network. As eventually all compounds of the network are produced, including all building blocks and cofactors, no partially conserved building blocks exist and 534 is an lower limit for the number of strictly conserved building blocks.

On the other hand, the number of nodes in the hierarchy graph calculated with cofactor functionalities (2098) gives the number of strictly and partially conserved building blocks. Hence, 2098 is an upper limit of the number of strictly conserved building blocks in the network.

The building blocks as defined here, are also accessable through calculation of the left side kernel of the stoichiometric matrix, as discussed in Schuster and Höfer [1991], Schuster and Hilgetag [1995], Imielinski et al. [2006]. This method yields weighted sets of metabolites which are conserved by all reactions in the network. Some of these sets represent strictly conserved building blocks (moieties in their nomenclature) which are present in the corresponding metabolites. The weights define the occurrence of the moieties within each metabolite. However, the method tends to produce an excessive number of conservation relations which makes it difficult to analyze for large metabolic networks. The scope hierarchy on the other hand indicates through its structure the most important conserved building blocks. The exact equivalence of the conserved entities predicted by the two methods still has to be shown in a later work.

5.4 The shape of metabolic networks

As discussed, the existence of building blocks and their conservation during all metabolic conversions strongly influences the shape and function of metabolic networks.

However, as argued before, not all reactions respecting this conservation rule are actually found in biological networks. The existence of metabolic reactions is further determined by various biological, thermodynamical and evolutionary factors. These may include the stability or toxicity of participating compounds, the velocity or directionality of reactions due to physiological limitations of the compound concentrations, the selection and development of enzymes suitable for a specific task and the avoidance of unnecessary compounds and reactions through evolutionary optimization.

While these factors have shaped the biological network as described in the KEGG database, the artificial metabolic network defined in this work lacks

such information. In order to compare the hierarchies of the two networks, in the artificial network the mentioned factors were approximated by a random selection of reactions from the set of all possible reactions.

As a result it turned out that major structural features of the KEGG hierarchy can be reproduced by the artificial network. This suggests that these structures are mainly determined by the conservation of building blocks. The observation of characteristic scopes representing specific element combinations in the KEGG hierarchy indicates that the atoms of many compounds can be relatively freely rearranged to form other compounds. This is in particular the case for the central parts of the network. Compounds in these parts can be synthesized from a small set of seed metabolites which essentially provide the chemical elements C, N, P, S as well as O and H to the network. This autotrophy of the central parts is a specific property of the analyzed KEGG network and cannot generally be expected from arbitrary metabolic networks.

The need for interconvertibility on the elementary level can be explained by the fact that the utilized KEGG network is an approximation of the metabolic capability of a whole ecosystem rather than of a single organism. Therefore, this network must be able to produce all its metabolites "from scratch" and cannot rely on other biological sources which could provide more complex substrates.

With the results from the studies of network modifications it can be concluded that evolution apparently has designed a network which has a sufficient set of reactions to allow interconvertibility between many compounds, while keeping the number of reactions small in order to avoid wasting resources for the production of an excessive number of enzymes.

On the other hand, the analysis of seeds reproducing the complete KEGG network indicates a quite large total number of building blocks (at least 534), inferring that the complete KEGG network is in fact not autotroph. One would assume that all more complex organic compounds used by biological organisms should also have been produced by biological metabolic systems. This should also be possible using the KEGG reference network which eventually should include metabolic data for all known organisms. Hence, apart from drugs, toxins and their degradation products which might have industrial origin, all metabolites in the network should be producible from a set of simple inorganic compounds with a size comparable to the number of chemical elements found in metabolism. As this is not the case, the KEGG database must miss various reactions responsible for the synthesis of the additional building blocks.

This result is actually not surprising. Even though the identification of metabolic pathways has been developed intensively during the last decades,

none of the investigated organism specific networks can be claimed to be complete. Furthermore, only a small fraction of all living organisms has been investigated at all. From that perspective, the number of compounds synthesizable from small inorganic seed compounds, i.e. more than 50% of all compounds in the network, seems comparably high. This indicates that a substantial part of metabolism is already covered by the KEGG database.

Also for many organism specific networks autotrophy cannot be assumed. Many heterotroph organisms require the uptake of more complex compounds which cannot be synthesized by their own networks. Such compounds, or more specifically the therein contained not synthesizable building blocks, necessarily have to be included in the seed. This assumption has been confirmed in a recent work where the nutritional requirements of various organisms have been predicted [Handorf et al., 2007].

Clearly, as the information on metabolic networks becomes more and more comprehensive, the KEGG network will further move to an autotroph limit, in which all compounds can be synthesized from small inorganic compounds. This does not mean that the corresponding hierarchy will only represent combinations of chemical elements as observed for a complete artificial network. In fact, partially conserved building blocks representing specific groups will also be allowed for autotroph networks and may indicate, as discussed before, as special role of these building blocks.

Organism specific networks on the other hand, in particular those of heterotrophs, will always require a larger number of building blocks including those produced by other members of an ecosystem. These ideas are sketched in figure 5.1.

The analysis of the effect of network modifications showed that the KEGG network is robust against the deletion of single and multiple reactions, indicating a large number of alternative synthesis paths. Clearly, as the KEGG network is a super set of all organisms in KEGG, it's robustness may simply result from alternative routes in different organisms. However, Ebenhöh et al. [2005] showed that also most organism specific networks show a robust behavior of the synthesizing capacity against reaction deletion.

This raises the more general question whether robustness against reaction removal is evolutionary favorable. The presence of a reaction is determined by the operation of a corresponding gene and a genetic defect may or may not lead to a inactivation of a reaction. Clearly, in case of a non-robust network, a mutation causing an inacivation of enzymes would led to the death of the offspring. However, this loss could be replaced by a healthy offspring. Also, all members of this non-robust species would have the advantage of a slim metabolism which does not require a lot of resources for its maintenance.

In contrast, having a more robust network increases the number of vital

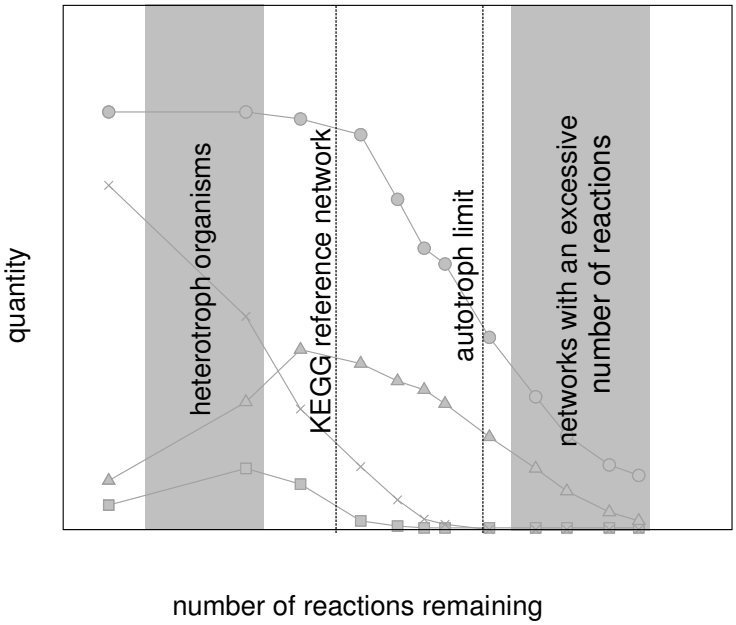


Figure 5.1: This sketch roughly places the discussed networks of heterotroph organisms, the current KEGG reference network and the putative autotroph limit in the context of artificial networks of different completeness. The curves are taken from figure 4.5 which represents the behaviour of the scope hierarchy of an artificial network when the number of reactions is changed.

but mutated descendants. The larger variability in the genome of the different descendants increases the probability that at least some individuals of the strain can survive substantial changes in the environment which is infact an evolutionary advantage.

Clearly, in an evolutionary optimal organism both factors should be considered. It can be expected that simple organism with fast reproduction rates would favor a slim design while more complex organism would prefer the robust behavior due to their higher costs of reproduction.

As discussed, the comparison of the KEGG network with a artificial network showed that the KEGG network does not have an excessive number of redundant reactions. On the other hand it has been found to be generally robust. The fact that there exist certain reactions whose removals are critical to the network function is conform with the above reasoning. While the general robustness of the network allows for genetic variability, individuals with a defect in a critical reactions will in fact die and be replaced by healthy relatives.

While the analyses in this work concentrated on the complete network,

similar results should be expected for organism specific networks. There, the number of critical reactions is assumably higher which however is no contradiction with the above reasoning.

The expansion process can, to a certain extend, also reflect the history of metabolic networks. While simple inorganic compounds like CO_2, NH_3, H_3PO_4 or H_2SO_4 can be assumed to be available even before the origins of life, more complex organic compounds require the existence of metabolic pathways for their synthesis. Clearly, the evolutionary development of enzymes catalyzing reactions whose substrates are not available in the environment seems unlikely. Therefore, it can be assumed, that new reactions have been selected by evolution if they could convert inorganic compounds or compounds produced by other metabolic reactions into useful products. Taking the set of current day metabolic reactions and inorganic compounds as seed, the method of network expansion defines a temporal order on the reactions which might relate to the actual historical development. Such an analysis has been introduced in an earlier paper [Ebenhöh et al., 2004]. Clearly, the exclusive utilization of present day reactions neglects the possibility that ancient reactions may have existed which were removed by evolutionary processes. Also, in metabolic networks many reactions occur in parallel and for those a temporal order cannot be inferred in the proposed manner. Hence, this topic leaves space for further investigation and should be addressed in a different work.

5.5 Conclusions

The purpose of this work is to analyze functional properties of metabolic networks. Previous investigations successfully analyzed the behavior of single metabolic pathways or even central parts of the networks of particular organisms. However, this breakdown into modules may miss important features which result from the interplay of the participating subunits. Such features are global properties of the metabolic networks and their investigation requires the analysis of the network as a whole.

Therefore, a fast and efficient method to determine functional capabilities of the networks is needed. The utilized method, the concept of scopes, fulfills this requirement. It proved in particular useful in uncovering and confirming biological principles just through a topological analysis of metabolism.

The analyses revealed design principles and evolutionary objectives behind the construction of current day metabolic networks, like the ability to synthesize its constituents from elementary building blocks (autotrophy), avoidance of superfluous reactions (slim design) on one side and allowance

of alternative reactions routes (robustness) on the other side. The analyzed KEGG network turned out to fulfill these objectives in particular for the well investigated central regions, assumably approaching an autotroph limit with increasing completeness. Despite of the current limitations of metabolic data, the results and principles presented in this work can be expected to remain valid when the metabolic knowledge further improves in future.

The concepts described can be applied to organism specific networks performing comparative analyses of their synthesizing capacities, nutritional requirements or robustnesses. Also, the functional capacities of the metabolism in different regulatory states can be analyzed by incorporating gene expression data. These topics have already been tackled in recent publications.

By refining the model of artificial metabolic networks, e.g. by considering thermodynamical or biological factors, further principles determining the shape of metabolic networks may be uncovered.

It will be useful to incorporate the metabolic knowledge from other biological databases. In turn, the methods presented here can be used to identify missing links in the databases, hence initiating an iterative process leading to a further improvement of metabolic knowledge. In this way this work opens a wide field for further investigation.

Appendix A

Additional Information

A.1 Method

The method defined in section 1.5 can be described by the flow chart depicted in figure A.1.

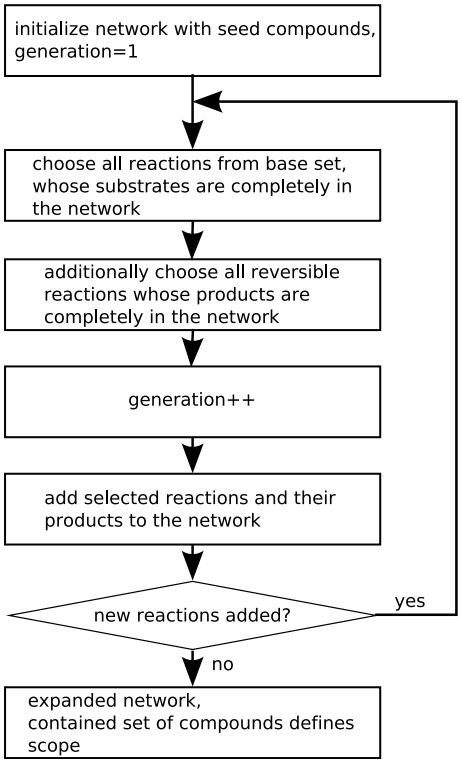


Figure A.1: The algorithm for the scope calculation.

As an example the following base set of reactions is considered:

$$C_1 + C_2 \longrightarrow C_3 + C_4 \tag{A.1}$$
$$C_3 + C_5 \longrightarrow C_6 + C_7 \tag{A.2}$$
$$C_2 + C_4 \longrightarrow C_8 \tag{A.3}$$

This reaction system can be represented as a graph as shown in figure A.2. Here, metabolic compounds are represented by circles while the arrows indicate the reactions between them. In figure A.2a the network is provided

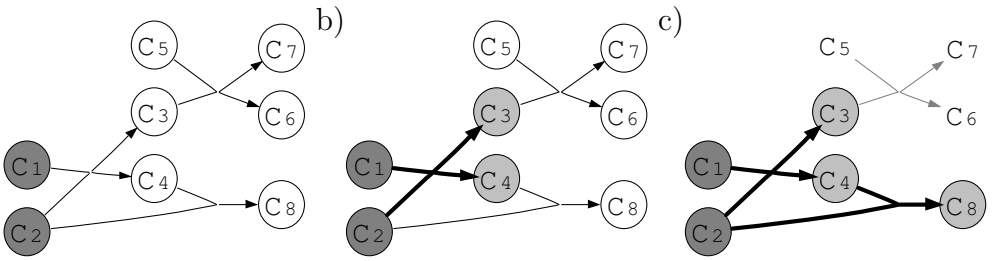


Figure A.2: Network expansion process with 3 generations with $\Sigma(C_1, C_2) = \{C_1, C_2, C_3, C_4, C_8\}$.

with the seed C_1, C_2 as indicated by the gray circles. In the first loop of the algorithm, only reaction A.1 can occur, producing the compounds C_3 and C_4. as shown in figure A.2b. Now, reaction A.3 can operate since C_4 has become available, producing C_8, cf. figure A.2c. Reaction A.2 cannot occur with the chosen seed since C_5 cannot be provided by the network. Consequently, the scope $\Sigma(C_1, C_2)$ is $\{C_1, C_2, C_3, C_4, C_8\}$.

A.2 Importing data from KEGG

For the calculations performed in this work the reactions have been taken form the KEGG database [Kanehisa, 1997, Kanehisa et al., 2006]. Specifically, the information about metabolic pathways is contained in the LIGAND part of that database. There exists several ways to access the data. Here, the information was imported from a text file representation of the database.

The relevant information was extracted from three files containing information on chemical compounds, reactions and enzymes. Figure A.3 gives an excerpt of the "compound" file. For each compound the file contains one entry. For this work only the information on the names and the formula were used. Figure A.4 shows a part of the "reaction" file. This file provides information on which compounds are converted into which product compounds and which enzymes can catalyze the reaction.

```
...

ENTRY       C00048                      Compound
NAME        Glyoxylate;
            Glyoxalate;
            Glyoxylic acid
FORMULA     C2H2O3
MASS        74.0004
REACTION    R00013 R00364 R00365 R00366 R00369 R00372 R00373 R00465
            R00466 R00468 R00469 R00470 R00471 R00472 R00473 R00474
            R00475 R00476 R00477 R00478 R00479 R00588 R00652 R00717
            R00776 R00932 R00933 R00934 R01180 R01957 R03040 R03121
            R03874 R05418 R05419 R05493 R05862 R05863
RPAIR       A00043 A00112 A00410 A00915 A00916 A00918 A00920 A00921
            A00923 A00925 A00927 A00929 A00931 A01182 A01183 A01345
            A02713 A02916 A03489 A05024 A05026 A05087 A06227 A06252
            A06785 A07357 A08910 A08949 A08966
PATHWAY     PATH: map00230  Purine metabolism
            PATH: map00260  Glycine, serine and threonine metabolism
            PATH: map00330  Arginine and proline metabolism
            PATH: map00627  1,4-Dichlorobenzene degradation
            PATH: map00630  Glyoxylate and dicarboxylate metabolism
            PATH: map00660  C5-Branched dibasic acid metabolism
ENZYME      1.1.1.26        1.1.1.29        1.1.1.79        1.1.3.15
            1.1.99.14       1.2.1.17        1.2.3.5         1.4.1.10
            1.4.2.1         1.4.3.3         1.4.3.19        2.2.1.5
            2.3.3.7         2.3.3.9         2.3.3.11        2.3.3.12
            2.6.1.4         2.6.1.35        2.6.1.44        2.6.1.45
            2.6.1.60        2.6.1.63        2.6.1.73        3.5.3.19
            4.1.1.3         4.1.1.47        4.1.2.14        4.1.3.1
            4.1.3.13        4.1.3.14        4.1.3.16        4.1.3.24
            4.3.2.3         4.3.2.5
DBLINKS     CAS: 298-12-4
            PubChem: 3350
            ChEBI: 16891
ATOM        5
            1   C6a C    -0.2241    0.1310
            2   C4a C     0.1483   -0.5207
            3   O6a O     0.1586    0.7793
            4   O6a O    -0.9793    0.1345
            5   O4a O     0.8966   -0.5241
BOND        4
            1     1   2 1
            2     1   3 1
            3     1   4 2
            4     2   5 2
///

...
```

Figure A.3: The content of the file "compound" describing chemical compounds in the metabolism. The entry for the compound C00048 is shown. The file contains such entries for each compound in the metabolism.

```
...

ENTRY       R00210                      Reaction
NAME        Pyruvate:NADP+ 2-oxidoreductase (CoA-acetylating)
DEFINITION  Pyruvate + CoA + NADP+ <=> Acetyl-CoA + CO2 + NADPH
EQUATION    C00022 + C00010 + C00006 <=> C00024 + C00011 + C00005
RPAIR       A00007 A05786
PATHWAY     PATH: rn00010  Glycolysis / Gluconeogenesis
            PATH: rn00620  Pyruvate metabolism
ENZYME      1.2.1.51        1.2.4.1         1.8.1.4         2.3.1.12
///

...
```

Figure A.4: The content of the file "reaction" describing chemical reactions in the metabolism. The entry for the reaction R00210 is shown. The file contains such entries for each reaction in the metabolism.

Figure A.5 shows a part of the "enzyme" file. This information is relevant if species specific networks are considered. In this file information can be found on whether an enzyme has a gene coding for it in a specific organism. Using the information about catalyzing enzymes from the "reaction" file one can determine, whether a certain reaction can occur in a specific organism or not.

For this work version 29a from 13th April 2005 of the LIGAND database was used comprising a total of 6401 reactions . During the import of the database certain curations have been applied to the data. First, the reactions have been checked for the conservation of chemical elements. 288 Reactions not fulfilling this condition have been excluded.

Second, 958 reactions dealing with compounds containing variable parts, e.g. $CHO_2R(CH_2)_n$ (long-chain carboxylate), have been excluded as the treatment of such compounds is difficult. In particular, if one reaction provides a variable compound and a second reaction requires a specific instance of that compound, these two compounds cannot easily be matched and the synthesis path is interrupted. The same occurs if the first reaction produces a unspecific compounds like "amino acid" while a second reaction metabolizes a specific one. Also, 344 reactions containing Glycan reactions have been removed as the analysis of Glycans a not a goal of this work. Such problems should be addressed in a later work.

Alltogether, a network of 4811 reactions and 4104 compounds was used in this analysis.

Apart from these more technical problems, the data is certainly incomplete. The uncovering of metabolic pathways is still an area of agile biological research. Therefore, it can be expected that the metabolic information in the KEGG database will be extended or corrected in the future.

```
...

ENTRY       EC 1.1.1.20
NAME        glucuronolactone reductase
            GRase
            gulonolactone dehydrogenase
CLASS       Oxidoreductases
            Acting on the CH-OH group of donors
            With NAD+ or NADP+ as acceptor
SYSNAME     L-gulono-1,4-lactone:NADP+ 1-oxidoreductase
REACTION    L-gulono-1,4-lactone + NADP+ = D-glucurono-3,6-lactone + NADPH + H+
SUBSTRATE   L-gulono-1,4-lactone
            NADP+
PRODUCT     D-glucurono-3,6-lactone
            NADPH
            H+
PATHWAY     PATH: map00053  Ascorbate and aldarate metabolism
GENES       BPS: BPSL2727(xdhB) BPSL2728(xdhA)
REFERENCE   1
            Suzuki, K., Mano, Y. and Shimazono, N. Conversion of
            L-gulonolactone to L-ascorbic acid; properties of the microsomal
            enzyme in rat liver. J. Biochem. (Tokyo) 48 (1960) 313-315.
DBLINKS     IUBMB Enzyme Nomenclature: 1.1.1.20
            ExPASy - ENZYME nomenclature database: 1.1.1.20
            ERGO genome analysis and discovery system: 1.1.1.20
            BRENDA, the Enzyme Database: 1.1.1.20
            CAS: 9028-30-2
///

...
```

Figure A.5: The content of the file "enzyme" describing chemical reactions in the metabolism. The entry for the enzyme EC 1.1.1.20 is shown. The file contains such entries for each enzyme in the metabolism.

These uncertainties in the underlying metabolism have been addressed in this work. It has been shown in chapter 4 that the general results of this work are persistent over large ranges of modifications in the network.

In figure A.6 it is shown how the scope size distribution is changed by updating the database to the version as of January 13, 2007. Clearly, even though the actual sizes are changed slightly, the general structure of the distribution of single and multi scopes remains the same.

Further, information on the reversibility of reactions has been extracted from the KGML files which specify the pathways for all organisms included in KEGG. In general, a particular reaction is listed in several KGML files and the information on its reversibility may be ambiguous. In fact, this is the case for 136 reactions. A reaction is considered to be irreversible only if it is defined as irreversible in all corresponding occurrences in the KGML files. This is the case for 2622 of the 5199 reactions.

For the prediction of cofactor pairs the information on structural overlaps between reactant pairs is used. For many reactant pairs this information can

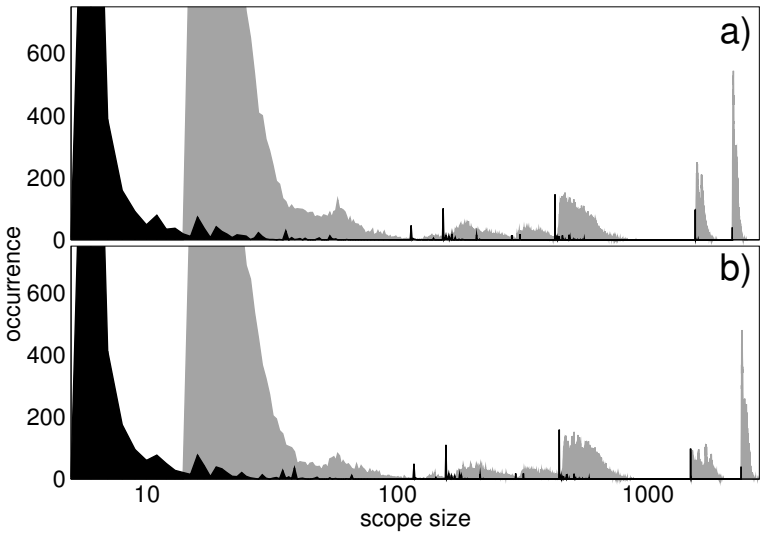


Figure A.6: The scope distributions of single (black) and 100000 random multi scopes (gray) for the KEGG database as of a) April 13, 2005 and b) January 13, 2007 are shown

be accessed in KEGG via Axxxx identifiers. The matching information can be found in the "ALIGN" section of the "rpair" file of the LIGAND database. This information has been used to compare the size of the overlapping part of two compounds with the sizes of the additional parts.

A.3 Modifications of the reaction network

This section describes certain technical modifications on the computer representation of the metabolic networks. This should not be confused with chapter 4 where the effect of possibly biologically inferred changes of the network are analyzed.

The first modification considers the abundance of water and its dissociation products oxygen and hydrogen. Water (KEGGID: C00001) is always added to the seed, unless otherwise stated. It can be biologically argued that for biochemical processes it is realistic to assume water to be always present. Due to dissociation this also holds, to a less extent, for oxygen and hydrogen. With the available reactions in the full KEGG network oxygen, hydrogen, H_2O_2 and O_2^- are automatically synthesized. Hence, the minimum scope size in this network is 5.

The presence of water may have a strong impact on particular scopes. For example, without water the scope size of APS is only 1 instead of 2183.

Without water the hydrolysis of APS to AMP and sulfate cannot be performed.

On the other hand, the general results of the large scale analysis performed in this work are mainly unaffected by the presence or absence of water. Figure A.7 shows the scope size distribution for the case with water and its dissociation products being present and the case where they are not explicitly added. Clearly, the number of single scopes resulting in large scopes is reduced. However, the positions of the characteristic scopes remain unchanged. For the scopes of random multi seeds also the frequencies are mainly unchanged. This can be explained by the fact, that when taking 15 arbitrary compounds, in most cases from at least one of them water can be produced.

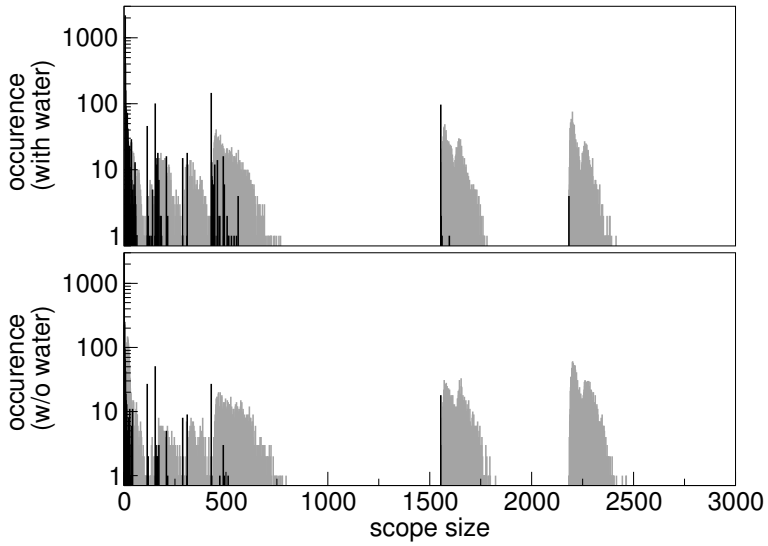


Figure A.7: The effect of water and its dissociation products on the scope sizes. The black curves indicate single scopes and gray curves 10000 scopes of 15 random seed compounds. The upper distribution shows scopes with water, etc. being always present, the lower graph shows scopes calculated without this modification. Note that the sets of random seeds are different for the two cases.

Analogous to water also other cofactors can be assumed as abundant. Clearly, if a biological cell converts external resources into the desired products it can rely on certain cofactors to be already present. In principle, in a growing and dividing cell also these cofactors have to be produced. For the analysis of smaller pathways, however, the abundance of cofactors may be a useful assumption. Furthermore, it should be noted that the production

of a cofactor may require the same cofactor to be present in first place, like for example ATP in glycolysis. Also in this case the assumption about the abundance of the cofactor is helpful.

Unlike in the case of water, it is not possible to simply add other cofactors. The reason is that this cofactor would be used to synthesize other compounds. In the case of ATP for example the minimum scope size would be 1554. Of course, the same arguments hold also for water, but the eventual effect is much smaller in this case since water only consists of the two element H and O.

In order to simulate only the functionality of a cofactor the following modifications were applied to the network: For cofactor pairs like ATP/ADP or NAD+/NADH, in all reactions containing the members of such a pair on different sides with the same stoichiometry, these cofactors were removed. The conservation of elements was corrected afterwards, i.e. for each ATP a phosphate and for each NADH a H^+ was added.

The cofactor coenzyme A had to be dealt with differently. Coenzyme A clips off acyl groups from one molecule and transfers them in a second reaction to a second molecule. In order to simulate its functionality it was added to the seed, but all reactions synthesizing or degrading it were manually removed.

Whereas water has always been assumed to be present, this is not true for other cofactors. Their influence is analyzed in more detail in section 2.5.

A.4 Derivation of the reversible Michaelis-Menten Equation

An enzymatic reaction

$$C \underset{k_-}{\overset{k_+}{\rightleftharpoons}} P$$

is split up into two sub reactions:

$$E + C \underset{k_{1-}}{\overset{k_{1+}}{\rightleftharpoons}} EC \underset{k_{2-}}{\overset{k_{2+}}{\rightleftharpoons}} E + P$$

and

$$\begin{aligned} v_1 &= k_{1+} e \prod_k c_k - k_{1-} z \\ v_2 &= k_{2+} z - k_{2-} e \prod_k p_k \end{aligned} \tag{A.4}$$

Here, C, P and E represent the substrates, products and the enzyme, respectively. the ks are the rates of the corresponding sub reactions and v_1 and v_2 are their effective velocities. z represents the concentration of the enzyme-substrate-complex, e the concentration of the free enzyme and c_k and p_k the concentrations of the substrates and products. If the two reactions proceed in a faster time scale than the changes in the metabolite concentrations of the substrates C and the products P, a quasi steady state approximation for z can be used:

$$\frac{dz}{dt} = v_1 - v_2 = 0.$$

$$\begin{aligned} 0 &= k_{1+} e \prod_k c_k - k_{1-} (\hat{e} - e) - k_{2+} (\hat{e} - e) + k_{2-} e \prod_k p_k \\ \hat{e} (k_{1-} + k_{2+}) &= e \left(k_{1+} \prod_k c_k + k_{1-} + k_{2+} + k_{2-} \prod_k p_k \right) \\ e &= \frac{\hat{e} (k_{1-} + k_{2+})}{k_{1+} \prod_k c_k + k_{1-} + k_{2+} + k_{2-} \prod_k p_k} \end{aligned}$$

with A.4

$$\begin{aligned} v_2 &= \hat{e} \left(k_{2+} - \frac{\left(k_{2+} + k_{2-} \prod_k p_k \right) (k_{1-} + k_{2+})}{k_{1+} \prod_k c_k + k_{1-} + k_{2+} + k_{2-} \prod_k p_k} \right) \\ &= \hat{e} \left(k_{2+} - \frac{\left(k_{2+} + k_{2-} \prod_k p_k \right)}{1 + \frac{k_{1+} \prod_k c_k}{k_{1-} + k_{2+}} + \frac{k_{2-} \prod_k p_k}{k_{1-} + k_{2+}}} \right) \end{aligned}$$

$$
\begin{aligned}
&= \hat{e}\left(\frac{k_{2+} - k_{2+} - k_{2-}\prod_k p_k + \frac{k_{1+}k_{2+}\prod_k c_k}{k_{1-}+k_{2+}} + \frac{k_{2-}k_{2+}\prod_k p_k}{k_{1-}+k_{2+}}}{1 + \frac{k_{1+}\prod_k c_k}{k_{1-}+k_{2+}} + \frac{k_{2-}\prod_k p_k}{k_{1-}+k_{2+}}}\right) \\
&= \hat{e}\left(\frac{\left(\frac{-k_{2-}k_{1-}\prod_k p_k + k_{1+}k_{2+}\prod_k c_k + k_{2-}k_{2+}\prod_k p_k - k_{2-}k_{2+}\prod_k p_k}{k_{1-}+k_{2+}}\right)}{1 + \frac{k_{1+}\prod_k c_k}{k_{1-}+k_{2+}} + \frac{k_{2-}\prod_k p_k}{k_{1-}+k_{2+}}}\right) \\
&= \hat{e}\left(\frac{\frac{k_{1+}k_{2+}\prod_k c_k}{k_{1-}+k_{2+}} - \frac{k_{2-}k_{1-}\prod_k p_k}{k_{1-}+k_{2+}}}{1 + \frac{k_{1+}\prod_k c_k}{k_{1-}+k_{2+}} + \frac{k_{2-}\prod_k p_k}{k_{1-}+k_{2+}}}\right)
\end{aligned}
$$

The reaction rate of the complete reaction v can be written as:

$$
\begin{aligned}
v = v_1 = v_2 &= \frac{\frac{V^+_{max}}{K^+}\prod_k c_k - \frac{V^-_{max}}{K^-}\prod_k p_k}{1 + \frac{\prod_k c_k}{K^+} + \frac{\prod_k p_k}{K^-}}, \\
V^+_{max} &= \hat{e}k_{2+}, \quad V^-_{max} = \hat{e}k_{1-} \\
K^+ &= \frac{k_{1-} + k_{2+}}{k_{1+}}, \quad K^- = \frac{k_{1-} + k_{2+}}{k_{2-}}
\end{aligned}
$$

A.5 Interconvertibilities

As mentioned in section 2.2, only compounds containing exactly the same elements can be interconvertible. However, the number of available reactions in metabolism is limited and thus not all such conversions exist. Within each group of compounds containing the same elements, all pairs of compounds were analyzed for being interconvertible. The corresponding numbers and the percentage of interconvertible pairs are given in table A.1. Within each group of compounds containing the same elements exists one or more sub groups whose compounds are interconvertible. If a compound is not interconvertible with any other compound, the corresponding sub group has a size of one. The last column in able A.1 gives the relative size distribution of these groups, where the shading is only used for distinguishing between the groups.

Elements	tot com	tot pair	num ic com	num ic pair	% ic pair	num groups	group dist
w/o	333	55278	41	29	0.0525%	18	
As.C.H.O	2	1	0	0	0%	0	
As.O	2	1	0	0	0%	0	
Br	1	1	0	0	0%	0	
Br.C.Cl.H	1	1	0	0	0%	0	
Br.C.H.N.O	3	3	0	0	0%	0	
Br.C.H.N.O.S	1	1	0	0	0%	0	
Br.C.H.O	5	10	4	6	60%	1	
Br.C.H.O.S	1	1	0	0	0%	0	
Br.H	1	1	0	0	0%	0	
C	1	1	0	0	0%	0	
C.Cl	1	1	0	0	0%	0	
C.Cl.H	25	300	12	13	4.33%	4	
C.Cl.H.N	6	15	2	1	6.67%	1	
C.Cl.H.N.O	15	105	5	4	3.81%	2	
C.Cl.H.N.O.P.S	2	1	0	0	0%	0	
C.Cl.H.O	71	2485	37	66	2.66%	11	
C.Cl.O	1	1	0	0	0%	0	
C.Co.H.N.O	13	78	2	1	1.28%	1	
C.Co.H.N.O.P	9	36	0	0	0%	0	
C.F.H.O	1	1	0	0	0%	0	
C.F.H.O.P	1	1	0	0	0%	0	
C.Fe.H.N.O	6	15	0	0	0%	0	
C.Fe.H.N.O.S	3	3	2	1	33.3%	1	
C.H	54	1431	7	5	0.349%	3	
C.H.I.N.O	4	6	0	0	0%	0	
C.H.I.O	5	10	0	0	0%	0	
C.H.Mg.N.O	10	45	7	11	24.4%	2	
C.H.N	62	1891	8	28	1.48%	1	
C.H.N.O	966	466095	382	9705	2.08%	88	
C.H.N.O.P	363	65703	161	4659	7.09%	22	
C.H.N.O.P.S	186	17205	89	214	1.24%	27	
C.H.N.O.P.Se	2	1	0	0	0%	0	
C.H.N.O.S	113	6328	37	123	1.94%	10	
C.H.N.O.Se	12	66	5	4	6.06%	2	
C.H.N.S	7	21	2	1	4.76%	1	
C.H.O	1501	1125750	575	6126	0.544%	153	
C.H.O.P	191	18145	94	1172	6.46%	14	
C.H.O.P.S	5	10	2	1	10%	1	
C.H.O.S	58	1653	3	3	0.181%	1	
C.H.O.X	1	1	0	0	0%	0	
C.H.S	3	3	0	0	0%	0	
C.H.Se	1	1	0	0	0%	0	
C.N	1	1	0	0	0%	0	
C.O	2	1	2	1	100%	1	
C.O.S	1	1	0	0	0%	0	
Cl	2	1	0	0	0%	0	
Cl.H	1	1	0	0	0%	0	
Cl.H.O	2	1	0	0	0%	0	

Elements	tot com	tot pair	num ic com	num ic pair	% ic pair	num groups	group dist
Co	1	1	0	0	0%	0	
F	1	1	0	0	0%	0	
Fe	1	1	0	0	0%	0	
H	1	1	0	0	0%	0	
H.N	2	1	2	1	100%	1	
H.N.O	5	10	3	3	30%	1	
H.N.O.P	1	1	0	0	0%	0	
H.O	3	3	2	1	33.3%	1	
H.O.P	5	10	4	6	60%	1	
H.O.P.Se	1	1	0	0	0%	0	
H.O.S	6	15	3	3	20%	1	
H.O.Se	1	1	0	0	0%	0	
H.S	1	1	0	0	0%	0	
H.Se	1	1	0	0	0%	0	
Hg	2	1	0	0	0%	0	
I	2	1	2	1	100%	1	
Mg	1	1	0	0	0%	0	
Mn	1	1	0	0	0%	0	
N	1	1	0	0	0%	0	
N.O	2	1	0	0	0%	0	
O	2	1	2	1	100%	1	
O.S	2	1	0	0	0%	0	
O.Se	1	1	0	0	0%	0	
S	1	1	0	0	0%	0	
X	1	1	0	0	0%	0	
all	4104	8419356	1568	23954	0.285%	387	

Table A.1:

Table A.1: Interconvertibilities of compounds in the KEGG network. The compounds are categorized according to their element content. The first column gives the element composition. "w/o" indicates all compounds without a formula given, "all" is the set of all compounds. The second column indicates the number of compounds and the third column the number of pairs with these compounds. The fourth column gives the number of compounds which are interconvertible with at least one other compound. The fifth column holds the number of interconvertible pairs. The sixth column gives the percentage of how many of all pairs are interconvertible. The seventh column gives the number of groups of interconvertible compounds which contain more than one compound. The last column shows the distribution of groups of interconvertible compounds. The shading is used to distinguish between neighboring groups.

A.6 Modelling of the expansion process

As discussed earlier, the general shape of an expansion curve can be explained by simple theoretical considerations. The change in the number of compounds is both proportional to the number of compounds so far discovered and proportional to the number of compounds still available from the eventual scope.

$$\dot{x} = x(1-x) = -x^2 + x \tag{A.5}$$

Here, the value is normalized to the total number of compounds in the scope. The differential equation can be solved by integration for a given initial value x_0, representing the seed.

$$t = \int_{x_0}^{x_t} \frac{1}{-x^2 + x} \tag{A.6}$$

$$x_t = \frac{e^t}{e^t - C_0} = \frac{1}{1 - C_0 e^{-t}} \tag{A.7}$$

The integration constant C_0 can be related to x_0 in the following way: $C_0 = \frac{x_0 - 1}{x_0}$. Even though it is a very rough approximation of the processes during the expansion, this analytical expression (cf. figure A.8) very well follows the general sigmoidal shape of the expansion process shown in figure 2.3.

Certainly, more sophisticated models for describing an expansion process can be developed. Such a model could for example consider that new reactions connect always to the new compounds of the last generation. Monomolecular reactions connect exclusively to the last generation

$$dy_{1,i} \propto dx_{i-1}, \tag{A.8}$$

whereas bi-molecular have at least one substrate in there

$$dy_{2,i} \propto x_{i-1} dx_{i-1} \tag{A.9}$$

and so on. Here, the $dy_{k,i}$ represent the numbers of new k-molecular reactions in the current generation i, whereas dx_{i-1} and x_{i-1} are the numbers of compounds in the last generation and in all previous generations, respectively. The number of new compounds dx_i would then be proportional to the number of new reactions:

$$dx_i \propto \sum_k \alpha_k y_{k,i}, \tag{A.10}$$

where the α_k weight the reaction types.

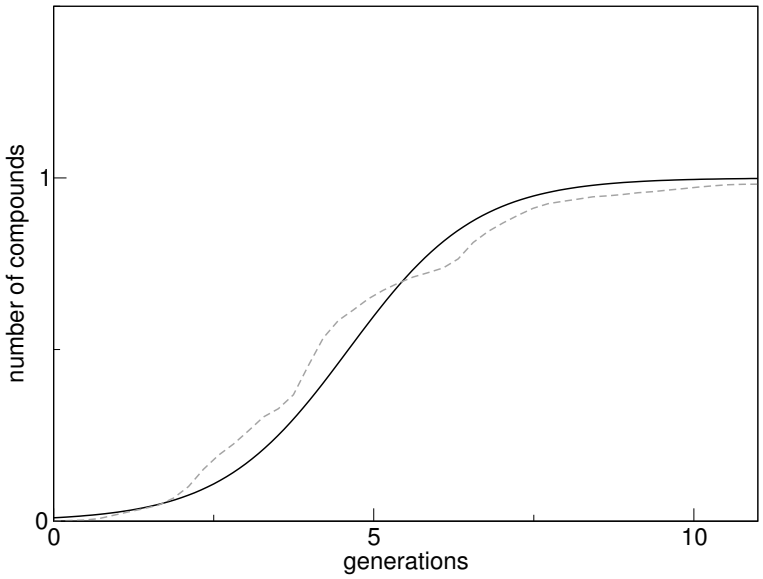


Figure A.8: Simple model of the expansion process (solid line): The number of compounds c in the expanded network is dependent on the generation t as follows: $c(t) = \frac{1}{1-C_0 e^{-t}}, C_0 = \frac{x_0-1}{x_0}$. Here, $x_0 = 0.01$ has been chosen. The dashed line shows the expansion process of APS (in arbitrary units).

A.7 Central metabolites and membrane transported metabolites

Table A.2 gives the detailed list of target metabolites defined as a minimal content of scopes which cover the central regions of metabolism. The set is obtained by determining all metabolites which are present in at least 90% of all organism specific networks in KEGG. See section 2.6 for more information.

It can be expected that this list of central metabolites as well as the metabolic networks defined in KEGG are not fully known yet. In particular, the absence of many amino acids is remarkable, since all organisms do protein synthesis. On the other hand, this absence may be explained by organisms which simply import these amino acids and directly use them in the translation process. In that case no metabolic reaction would be involved.

This list of central metabolites is similar to the definition of biomass, as for example used in Edwards et al. [2001]. The biomass in principle describes a set of metabolites which are necessary for cell division, i.e. metabolites which are essential for rebuilding a whole cell. This set usually contains amino acids, cell wall constituents, nucleotides for DNA and RNA, important cofactors and energy equivalents like sugars or ATP. As cell division is most

Acetyl-CoA	FADH2	L-Valine
Adenosine	Formate	NAD+
ADP	GDP	NADH
AMP	Glycerone phosphate	NADP+
ATP	Glycine	NADPH
beta-D-Fructose 6-phosphate	GMP	NH3
CDP	GTP	Nicotinate D-ribonucleotide
CMP	H+	Orthophosphate
CO2	H2O	Oxaloacetate
CoA	H2O2	Oxygen
CTP	IMP	Phosphoenolpyruvate
D-Erythrose 4-phosphate	Isopentenyl diphosphate	Pyrophosphate
D-Fructose 6-phosphate	L-Alanine	Pyruvate
D-Glyceraldehyde	L-Arginine	S-Adenosyl-L-homocysteine
D-Ribose 5-phosphate	L-Asparagine	S-Adenosyl-L-methionine
D-Ribulose 5-phosphate	L-Aspartate	sn-Glycerol 3-phosphate
dADP	L-Cysteine	Tetrahydrofolate
dAMP	L-Glutamate	UDP
dATP	L-Glutamine	UMP
dCDP	L-Histidine	UTP
dCTP	L-Isoleucine	Xanthosine 5'-phosphate
dGDP	L-Leucine	(2R)-2-Hydroxy-3-(phosphonooxy)-propanal
dGMP	L-Lysine	2,3-Bisphospho-D-glycerate
dGTP	L-Methionine	2-Phospho-D-glycerate
dTDP	L-Ornithine	3-Dehydroquinate
dTMP	L-Phenylalanine	3-Dehydroshikimate
dTTP	L-Proline	3-Phospho-D-glycerate
dUDP	L-Serine	3-Phospho-D-glyceroyl phosphate
dUMP	L-Threonine	5,10-Methenyltetrahydrofolate
dUTP	L-Tryptophan	5,10-Methylenetetrahydrofolate
FAD	L-Tyrosine	5-Phospho-alpha-D-ribose 1-diphosphate

Table A.2: List of target metabolites. These 93 compounds, which can be found in over 90% of all organsims in KEGG, are considered as central. The list ist sorted alphabetically.

essential for all organisms, the similarity to the metabolites in table A.2 can be expected.

It is useful to use compounds as seeds which are known to be able to pass the cell wall. Using the compounds defined by the KEGG database as substrates to ABC or PTS transporters will give reasonable candidates for such exchangeable compounds. As certain metabolites might be missing, the algorithm described in section 2.6 is able to use other compounds as seeds as well, but will prefer those given here (table A.3).

A.8 Calculation of synthesis paths

This section describes the algorithm for the calculation of synthesis paths from certain start metabolites (seed) to a target metabolite as used in section 2.7.

Clearly, when starting a network expansion from the seed and the target

ABC transported (PATH02010)	PTS transported (PATH02060)	small inorganic compounds
Betaine	alpha,alpha-Trehalose	Acetamide
Butyro-betaine	alpha-D-Glucose	Acetate
Capsular polysaccharide	Arbutin	Allyl alcohol
Carnitine	Ascorbate	Carbamate
Choline	beta-D-Glucose	Carbonic acid
Choline sulfate	beta-D-Glucoside	Chloride
Cobalt	Cellobiose	Cl-
Crotono-betaine	D-Fructose	CO2
Cyclomaltodextrin	D-Glucosamine	Cobalt
D-Allose	D-Glucose	Dimethylamine
D-Aspartate	D-Sorbitol	Ethanol
D-Galactose	Galactitol	Ethanolamine
D-Glucose	Glucose	Ethylamine
D-Methionine	Lactose	Fe2+
D-Ribose	Maltose	Fe3+
D-Xylose	Mannitol	Formate
Fe(III)dicitrate	N-Acetyl-D-glucosamine	Glycine
Fe(III)hydroxamate	N-Acetylgalactosamine	Glycolate
Fe-enterobactin	Nitrogen	H+
Fe2+	Salicin	H2O
Fe3+	Sorbose	HCO3-
Ferrichrome	Sucrose	HO-
Heme		Imidazole
Hemine		Iron
Iron chelate		Magnesium
L-Arabinose		Manganese
L-Arginine		Methane
L-Aspartate		Methanol
L-Glutamate		Methylguanidine
L-Glutamine		NH3
L-Histidine		Nitrate
L-Isoleucine		Nitric oxide
L-Leucine		Nitrite
L-Lysine		Nitrogen
L-Methionine		Nitrous oxide
L-Ornithine		Oxygen
L-Proline		Propan-2-ol
L-Threonine		Propane-1-ol
L-Valine		Propanoate
Lipo-oligosaccharide		Sulfur
Lipopolysaccharide		Trimethylamine N-oxide
Lipoprotein		Urea
Maltose		(R)-1-Aminopropan-2-ol
Manganese		1,3-Diaminopropane
Molybdate		1-Aminopropan-2-ol
Nickel		1-Butanol
Nitrate		
Orthophosphate		
Putrescine		
sn-Glycerol 3-phosphate		
Sodium		
Spermidine		
Sulfate		
Taurine		
Teichoic acid		
Tetrabenazine		
Thiamin		
Thiosulfate		
Tungsten		
Urea		
Vitamin B12		
Zinc		
2,6-Dimethoxybenzoquinone		
2-(beta-D-Glucosyl)-sn-glycerol		

Table A.3: List of transported metabolites. Given are metabolites transported by ABC transporters as specified in the KEGG pathway PATH02010 and metabolites transported by the phosphotransferase system as in PATH02060 and small inorganic compounds which might be sources for autotrophic organisms.

compound can be reached, one or more possible paths will be subsets of the expanded network. In order to get paths which contain only reactions which are necessary for the specified conversion, all other reactions including possible redundant reactions must be filtered out.

Therefore, for each reaction it is memorized which reactions provided the necessary substrates for the incorporation during the expansion process. It should be noted that only those reactions are memorized which produce a compound in the step of incorporation of that compound. Reaction additionally producing a compound in a later step are ignored. Occasionally several reactions produce the same compound in the same step which eventually leads to alternative paths.

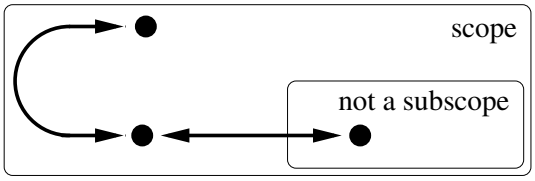
In order to determine a possible synthesis path a backtracking is started from the target compound. In each step a reaction providing one of the necessary compounds is selected. If there exists several reactions producing the same compound, as described above, one is chosen. Clearly the algorithm branches if there exist more than one substrate for a reaction. The algorithm ends if all followed branches reach one of the seed compounds.

The algorithm will find a synthesis path which is minimal in the number of subsequent synthesis steps. As there may exist parallel branches the minimality of the total number of reaction is not assured. Also, there may exists paths which have a larger number of subsequent steps which hence cannot be found by the algorithm.

A.9 Existence of single subscopes

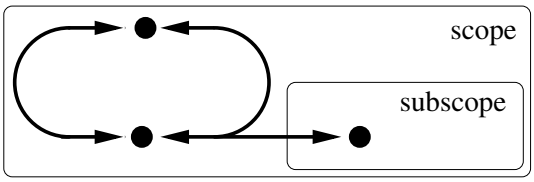
As shown in Figure 3.2a in section 3.2 there exists only four nodes in the hierarchy of single scopes which have an out degree of exactly one. This can be explained as follows:

Generally, if reactions are reversible, each scope which is not a sink in the hierarchy has at least two sub scopes. The reason is that a reaction generating compounds that are not interconvertible with the seed needs to generate at least two not interconvertible products. Otherwise, this reaction could be used in the opposite direction using the products as seed:



There exist, however, a rare situation where a scope may have exactly one subscope: If a reaction produces a product that is a seed compound of the

superscope and another that is not, then the superscope has one successor which alone cannot produce the compounds of the superscope:



This situation occurs four times in the network. One example is the scope of Hydrogen sulfide (H_2S) which is a subscope of the scope of Sulfite/Sulfur. The two reactions

$$O_2 + H_2O + H_2 + 2\text{Sulfur} \rightleftharpoons H_2S + \text{Sulfite}$$
$$\text{Sulfur} + O_2 + H_2O \rightleftharpoons \text{Sulfite}$$

generate Sulfite from Sulfur, one of them additionally produces Hydrogen sulfide. Sulfite is interconvertible with Sulfur (since H_2O and O_2 are present. From H_2S the two other compounds cannot be synthesized.

A.10 Artificial networks

The artificial networks used in this work consist of reactions which perform transformations between artificial compounds. These compounds are represented by sets of building blocks. In a specific network there exists a finite set B of building blocks. Each compound c_j can contain at most N_i units of the building block i and is uniquely defined by the set of building blocks and the corresponding numbers of units $b_{ij}, i \in B$. Hence the number of possible compounds c in the network is

$$c = \left(\prod_{i}^{B}(N_i + 1)\right) - 1. \tag{A.11}$$

Reactions are required to conserve the building blocks, i.e. the number of units of any building block i in the substrates must be the same as the number of units of the building block in the products. For a reaction with the set of substrates Q and the set of products P this means

$$\sum_{k \in Q} -s_k b_{ik} = \sum_{l \in P} s_l b_{il}, \forall i \in B, \tag{A.12}$$

with b_{ij} is the number of building blocks of type i in compound j and s_j being its stoichiometry in this reaction. (Note: for compounds occurring on

both sides of a reaction, the stoichiometries can be modified in such a way that this compounds remains on only one side)

The number of all possible reactions between the above defined compounds obeying the conservation relation is infinite. Clearly, the set of substrates can be chosen from the 2^c possible sets of compounds. Still, for each compound in this set the stoichiometry s_j can be freely chosen.

However, reactions of type $q \leftrightarrow p$, with q being the number of substrates and p the number of products and $q, p > 1$, can be replaced by a set of reactions of type $2 \leftrightarrow 1$: A reaction of type $q \rightarrow p$ transfers from each of their q substrates a set of building blocks to each of their p products. Some of these sets may be empty. This splitting-up of the substrates can be performed by at most $q \times (p-1)$ reactions of type $1 \rightarrow 2$. Each of the p products receives a set of building blocks from each of the q substrates. Again, some of these sets may be empty. This assembly of the products can be performed by at most $p \times (q-1)$ reactions of type $2 \rightarrow 1$. By utilizing the split-up-reactions before the assembly-reactions it can be assured that the intermediates do not violate any size limitations on the compounds as long as the substrates and products obey them.

Compounds with stoichiometries s_j larger than one can be treated as s_j separate compounds. Hence, with the above substitution, reactions of type $q \leftrightarrow p$ with arbitrary stoichiometries can be represented as a set of reactions of type $2 \leftrightarrow 1$ with all stoichiometries being 1.

Reactions of type $1 \leftrightarrow 1$ with stoichiometries equal to 1 do not play a role for these networks as the construction rule for the compounds does not allow for any distinguishable isomers. $1 \leftrightarrow 1$ reactions with larger stoichiometries can be accordingly transformed. Consequently, the set of $2 \leftrightarrow 1$ reactions contains as a special case reactions of type $2c_1 \rightleftharpoons c_2$.

The number of $2 \leftrightarrow 1$ reactions r with unity stoichiometries in a network with size limited compounds, as defined above, is finite. The number of reactions r can be expressed as follows:

$$r = \sum_{j=1}^{c} R(c_j), \tag{A.13}$$

where $R(c_j)$ is the number of possible reactions that can split up c_j into c_u and c_v:

$$c_j \rightleftharpoons c_u + c_v \tag{A.14}$$

$R(c_j)$ is dependent on the number of units of a specific building block i in the compound j, b_{ij}. With equation A.12 it follows:

$$b_{ij} = b_{iu} + b_{iv}, \forall i \in B. \tag{A.15}$$

Apparently, any set of the b_{iu} with $0 \leq b_{iu} \leq b_{ij}$ yields a possible split-up-reaction of c_j as in equation A.14. The compound c_v is then defined by $b_{iv} = b_{ij} - b_{iu}$. The two cases where $b_{iu} = 0$ for all $i \in B$ and $b_{iv} = 0$ for all $i \in B$ will be excluded, as the corresponding reactions do not perform a transformation. The number of possible sets of the b_{iu} is

$$\left[\prod_{i \in B} (b_{ij} + 1)\right] - 2 \tag{A.16}$$

This is, however, not the number of possible split-up-reactions. Clearly, swapping c_u and c_v yields the same reaction. Hence, except for the case where $b_{iu} = b_{iv}$ for all $i \in B$ each reaction is generated twice. The case $b_{iu} = b_{iv}$ for all $i \in B$ can only occur if all b_{ij} are even. Hence, $R(c_j)$ can be expressed as follows:

$$R(c_j) = \left(\frac{1}{2}\prod_{i \in B} b_{ij} + 1\right) - 1 + g(b_{ij}) \ , \quad g(b_{ij}) = \begin{cases} \frac{1}{2} & \text{, if all } b_{ij} \text{ even} \\ 0 & \text{, otherwise} \end{cases} \tag{A.17}$$

This expression can be inserted into equation A.13. As the c_j are defined by the b_{ij} the sum over all compounds can be replaced by a sum over all combinations of the b_{ij}. The index j numerating the compounds disappears, as the b_i are now enumerated directly:

$$r = \sum_{b_1=0}^{N_1} \cdots \sum_{b_{|B|}=0}^{N_{|B|}} \left[\left(\frac{1}{2}\prod_{i=1}^{|B|} b_i + 1\right) - 1 + g(b_1, \ldots, b_{|B|})\right] \tag{A.18}$$

$$r = \frac{1}{2}\left(\sum_{b_1=0}^{N_1} (b_1 + 1)\right) \cdot \ldots \cdot \left(\sum_{b_{|B|}=0}^{N_{|B|}} \left(b_{|B|} + 1\right)\right) - \left(\sum_{b_1=0}^{N_1} 1\right) \cdot \ldots \cdot \left(\sum_{b_{|B|}=0}^{N_{|B|}} 1\right) + \frac{1}{2} n_{even} \tag{A.19}$$

Here, $|B|$ is the number of building block types. n_{even} is the number of combinations of the b_i for which all b_i are even, including the case $b_i = 0$ for all $i \in B$. Equation A.19 can be rewritten as:

$$r = \left[\left(\frac{1}{2}\right)^{(|B|+1)} \prod_{i=1}^{|B|} (N_i + 1)(N_i + 2)\right] - \prod_{i=1}^{|B|} (N_i + 1) + \frac{1}{2}\prod_{i=1}^{|B|} (N_i \text{ div } 2 + 1), \tag{A.20}$$

where N_i div 2 denotes the largest integer smaller than or equal to $N_i/2$. The 3rd term in equation A.20 is by about a factor of $2^{|B|+1}$ smaller than the

2nd term. Hence, for a sufficiently large number of building block types $|B|$, $g(b_i)$ may be neglected:

$$r \approx \left[\left(\frac{1}{2}\right)^{(|B|+1)} \prod_{i=1}^{|B|}(N_i+1)(N_i+2)\right] - \prod_{i=1}^{|B|}(N_i+1) \tag{A.21}$$

The 2nd term in equation A.20 is approximately the number of possible compounds c in the network (cf. equation A.11). For sufficiently large N_i, the number of possible reactions r is much larger than the number of compounds c and also this term can be neglected:

$$r \approx \left(\frac{1}{2}\right)^{(|B|+1)} \prod_{i=1}^{|B|}(N_i+1)(N_i+2) \tag{A.22}$$

Even though it is not a problem to calculate equation A.20, it should be noted that the approximations A.21 and A.22 give already good results for the networks used in this work. For the network defined by $N_{(A,B,C,D,E)} = (6,4,4,3,2)$, (A.20) yields 186972, (A.21) 186900 and (A.22) 189000 possible reactions. For the case that the $N_i \approx N$ from equation A.11 and A.22 can be seen that the number of possible compounds in a network is of the order of $N^{|B|}$ and the number of possible reaction of the order of $N^{2|B|}/2^{|B|+1}$.

A.11 The hierarchy graph

Generally, a graph G can be represented by its adjacency matrix g_{ij} which is defined as follows:

$$g_{ij} = \begin{cases} 1\ , & \text{if there exists a directed edge from node } i \text{ to node } j \\ 0\ , & \text{otherwise} \end{cases} . \tag{A.23}$$

The scope hierarchy graph S is based on the scope inclusion graph K which is defined as follows:

$$k_{ij} = \begin{cases} 0\ , & \text{if } \Sigma_j \not\subset \Sigma_i \\ 1\ , & \text{if } \Sigma_j \subset \Sigma_i \end{cases} , \tag{A.24}$$

The Σ_u can be arbitrary sets of scopes in the network, for example the set of single scopes $\Sigma(c_u)$ of all compounds c_u in the network. In that case it should be noted that a group of interconvertible compounds c_v is represented by only scope (cf. equation 1.18) and hence by only one node in the graph. As there exist no two scopes Σ_u and Σ_v for which $\Sigma_u \subset \Sigma_v \wedge \Sigma_v \subset \Sigma_u \wedge \Sigma_u \neq \Sigma_v$ is true, the graph K is acyclic.

The scope hierarchy S can be derived from K by removing redundant edges between any node i and j if there exists at least one node l for which $\Sigma_j \subset \Sigma_l \subset \Sigma_i$ holds:

$$s_{ij} = \begin{cases} 0\ , & \text{if } (k_{ij} = 0) \text{ or } (\exists l \text{ with } k_{il} = 1 \wedge k_{lj} = 1) \\ 1\ , & \text{otherwise} \end{cases} , \tag{A.25}$$

Following the idea of the building blocks, a scope inclusion $\Sigma_l \subset \Sigma_i$ means that all compounds together in Σ_l contain generally less building block types (and definitely not more) than all compounds together in Σ_i. For a further scope Σ_j which is a subscope of Σ_l, $\Sigma_j \subset \Sigma_l$, clearly also the relation $\Sigma_j \subset \Sigma_i$ holds. The corresponding directed edge between nodes i and j in graph K, however, conceals that Σ_j contains an even smaller subset of building blocks than Σ_l. Removing this edge in S still leaves the path $i \rightarrow l \rightarrow j$ which describes the inclusion relation between i and j and is a much more appropriate description of the building block distribution in these scopes.

A.12 Graph layout

For the graph visualizations presented in this work the Graphviz package (Gansner and North [2000]) has been used. In particular, the layouter dot, producing hierarchical layouts, has been utilized for the hierarchies and the synthesis paths.

Hierarchical layout basically aligns directed edges to a preferred direction (e.g. top to bottom). In doing so, a certain order is applied to the nodes, positioning sources rather to the top and sinks to the bottom of the layout. More details can be found in the above cited reference.

The visualization of the hierarchy graph introduced in chapter 3 with dot is straight forward. Since the hierarchy constitutes as directed acyclic graph the hierarchical layout ensures that nodes representing sub scopes are displayed beneath their corresponding super scope nodes.

Also the synthesis paths in section 2.7 were layouted by dot. However, the corresponding reaction set first has to be converted in a suitable graph representation. Here, the bi-partite graph representation introduced in section 1.2 is used, where directed edges are used pointing from substrates to the corresponding reactions and from reactions to their corresponding products. The direction of a reaction, i.e. the definition of substrates and products is determined by the direction in which the reactions is used in the synthesis path to be displayed.

This graph representation can be used as input for the hierarchical layouter dot. In doing so, it is possible to follow the synthesis steps intuitively

in a preferred direction.

Furthermore, the same layouter can be used to visualize arbitrary metabolic networks. Here, cycles may occur. Still, the hierarchical layouter dot can handle such graphs, resulting in graphs where edges may point in the opposite of the preferred direction. Additionally, the graph representation of reaction sets can be further modified for clearer layout. For example, for larger networks is is useful to repeat nodes for highly utilized substances in order to avoid long and potentially crossing edges.

A.13 Non expanding double scopes are unique

Two distinct pairs of seed compounds (A,B) and (C,D) are considered. Here distinct means that at least one compound of the second pair is not interconvertible with both compounds of the first pair, i.e. without loss of generality $\Sigma(C) \neq \Sigma(A)$ and $\Sigma(C) \neq \Sigma(B)$. May the unions of the single scopes of the two compounds in each pair be identical, i.e. $\Sigma(A) \cup \Sigma(B) = \Sigma(C) \cup \Sigma(D)$?

It is clear, that if $\Sigma(A)$ and $\Sigma(B)$ are disjoint then there exists no such $\Sigma(C)$ and $\Sigma(D)$.

If $\Sigma(A)$ and $\Sigma(B)$ are not disjoint, then without loss of generality, C is in $\Sigma(A)$. Then $\Sigma(C)$ must be smaller than $\Sigma(A)$, if A and C are not interconvertible. Hence A is not in $\Sigma(C)$. Thus it must hold: $A \in \Sigma(D)$ and $\Sigma(B) \subset \Sigma(D)$. This is only possible if either $\Sigma(A) \subset \Sigma(B)$ or $\Sigma(B) \subset \Sigma(A)$. In that case $\Sigma(D)$ must be identical to the larger scope.

Hence, non-expanding double scopes, i.e. $\Sigma(A, B) = \Sigma(A) \cup \Sigma(B)$ are unique, unless one of the single scopes is included in the other.

Of the 4149610 pairs resulting in non-expanding scopes only 18564 pairs show such an inclusion. It is also clear that in these cases the resulting double scope is identical to the larger single scope.

A.14 Reduction of the total number of scopes by single reactions

A reaction R shall posses n substrates and m products and shall further be reversible. Without R, the total number of scopes as defined by the $n + m$ compounds is 2^{n+m}, as the $n + m$ compounds are isolated and hence each possible set of these compounds is a scope. With reaction R, all seeds containing either all n substrates or all m products expand to the full set of $n + m$ compounds. If all substrates are present there exist 2^m possibilities to choose the products and if all products are present 2^n possibilities for the

substrates are possible. Altogether, there exist $2^n + 2^m - 1$ seeds which yield as scope the full set of $n + m$ compounds. All other seed combinations lead to distinguished scopes. Therefore, there exist $2^{n+m} - (2^n + 2^m - 1) + 1$ scopes if R is present. Dividing by the total number of scopes without R one yields:

$$r = 1 - \frac{1}{2^m} - \frac{1}{2^n} + \frac{1}{2^{n+m-1}}. \tag{A.26}$$

If a network consists of several isolated parts, the total number of scopes is the product of the number of scopes in each part. Clearly, in a network with N compounds and no reactions has 2^N possible scopes. Each reaction which is put in the network and which is not connected to any other reaction will reduce the number of scopes by the above given factor r. For the typical case of an $2 \leftrightarrow 2$ reaction, this factor becomes $r = 0.625$, for a $1 \leftrightarrow 1$ reaction $r = 0.5$.

It is in fact possible to place 339 isolated reactions in the KEGG network. Using equation A.26, the upper limit for the total number of scopes in the KEGG network can be reduced by a factor of approximately 2^{320}.

Clearly, also the addition of connected reactions will decrease the number of scopes. The quantitative determination of the effect is however difficult and dependent on the topology of the whole network.

A.15 Software tools

In order to perform the calculations and visualizations presented in this work, the following software tools were used:

Calculations:

- Perl
- C (scope routines)
- PDL (perl package)

Visualization:

- Graphviz (Gansner and North [2000])
- Grace (Team)
- Matlab

An online demo of the algorithms is available (Handorf and Ebenhöh [2007]).

Bibliography

M. Arita. The metabolic world of Escherichia coli is not small. *Proceedings of the National Academy of Sciences*, 101(6):1543–1547, 2004. doi: 10.1073/pnas.0306458101. URL `http://www.pnas.org/cgi/content/abstract/101/6/1543`.

Peter W. Atkins. *Physikalische Chemie*. VCH Verlagsgesellschaft, Winheim, Germany, 1990. ISBN 3-527-25913-9.

A. Bairoch. The enzyme database in 2000. *Nucleic Acids Res.*, 28:304–305, 2000.

A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

H. P. J. Bonarius, G. Schmid, and J. Tramper. Flux analysis of underdetermined metabolic networks: The quest for the missing constraints. *Trends Biotech.*, 15:308–314, 1997.

N. Christian, T. Handorf, and O. Ebenhöh. Metabolic synergy: Increasing biosynthetic capabilities by network cooperation. *Genome informatics*, 2007. in print.

B. L. Clarke. Complete set of steady states for the general stoichiometric dynamical system. *Journal of Chemical Physics*, 75:4970–4979, November 1981.

P. Dittrich and P.S. di Fenizio. Chemical organisation theory. *Bull Math Biol.*, 69(4):1199–1231, 2007.

O. Ebenhöh and W. Liebermeister. Structural analysis of expressed metabolic subnetworks. *Genome Informatics*, 17(1):163–172, 2006.

O. Ebenhöh, T. Handorf, and R. Heinrich. Structural analysis of expanding metabolic networks. *Genome Informatics*, 15:35–45, 2004.

O. Ebenhöh, T. Handorf, and R. Heinrich. A cross species comparison of metabolic network functions. *Genome Informatics*, 16(1):203–213, 2005.

O. Ebenhöh, T. Handorf, and D. Kahn. Evolutionary changes of metabolic networks and their biosynthetic capacities. *IEE Proc. Syst. Biol.*, 153(5): 354–358, 2006.

J. S. Edwards, R. U. Ibarra, and B.O. Palsson. In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. *Nature Biotechnology*, 19:125–130, 2001.

J.S. Edwards and B.O. Palsson. The escherichia coli mg1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proc Natl Acad Sci USA*, 97:5528–5533, 2000.

W. Fontana and L.W. Buss. The arrival of the fittest: toward a theory of biological organization. *Bulletin of Mathematical Biology*, 56(1):1–64, 1994.

M. Y. Galperin. The molecular biology database collection: 2006 update. *Nucleic Acids Res.*, 34:D3–D5, 2006.

E. R. Gansner and S. C. North. An open graph visualization system and its applications to software engineering. *Software — Practice and Experience*, 30(11):1203–1233, 2000. URL `http://citeseer.ist.psu.edu/gansner99open.html`.

H. Genrich, R. Küffner, and K. Voss. Executable petri net models for the analysis of metabolic pathways. *Int J STTT*, 3:394–404, 2001.

T. Handorf and O. Ebenhöh. Minimal sets of external compounds in metabolic networks. *Genome Informatics IBSB Poster Abstracts*, pages 11–12, 2004. URL `http://www.jsbi.org/journal/IBSB04/IBSB04P005.pdf`.

T. Handorf and O. Ebenhöh. Metapath online: a web server implementation of the network expansion algorithm,. *Nucleic Acids Research*, 35(webserver issue):W613–W618, 2007. URL `http://scopes.biologie.hu-berlin.de`.

T. Handorf, O. Ebenhöh, and R. Heinrich. Expanding metabolic networks: Scopes of compounds, robustness, and evolution. *J. Mol. Evol.*, 61:498–512, 2005.

T. Handorf, O. Ebenhöh, D. Kahn, and R. Heinrich. Hierarchy of metabolic compounds based on their synthesizing capacity. *IEE Proc. Systems Biology*, 153(5):359–363, 2006.

T. Handorf, O. Ebenhöh, N. Christian, and D. Kahn. An environmental perspective on metabolism,. *Journal of Theoretical Biology*, 2007. in print.

R. Heinrich and T.A. Rapoport. A linear steady-state treatment of enzymatic chains. *European Journal of Biochemistry*, 42(1):97–105, 1974.

R. Heinrich and S. Schuster. *The regulation of cellular systems*. Chapman & Hall, New York, 1996. ISBN 0-412-03261-9.

R. Heinrich and S. Schuster. The modelling of metabolic systems. structure, control and optimality. *Biosystems*, 47(1-2):61–77, 1998.

N.H. Horowitz. On the evolution of biochemical syntheses. *PNAS*, 31:153–157, 1945.

M. Imielinski, C. Belta, H. Rubin, and A. Halasz. Systematic analysis of conservation relations in *Escherichia coli* genome-scale metabolic network reveals novel growth media. *Biophys. J.*, 90:2659–2672, 2006.

N. Jamshidi, J.S. Edwards, T. Fahland, G.M. Church, and B.O. Palsson. Dynamic simulation of the human red blood cell metabolic network. *Bioinformatics*, 17(3):286–287, 2001.

H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

M. Kanehisa. A database for post-genome analysis. *Trends Genet.*, 13:375–376, 1997.

M. Kanehisa, S. Goto, M. Hattori, K.F. Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, 34:D354–357, 2006.

P.D. Karp, C.A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin, and N. Lopez-Bigas. Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, 19:6083–6089, 2005.

S.A. Kauffman. Autocatalytic sets of proteins. *Journal of Theoretical Biol.*, 119(1):1–24, 1986.

I. Koch, B.H. Junker, and M. Heiner. Application of petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics*, 21(7):1219–1226, 2005.

W. Martin and M.J. Russell. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Phil. Trans. R. Soc. Lond. B*, 358:59–85, 2003.

I. Oancea and S. Schuster. Topological analysis of metabolic networks based on petri net theory. *In Silico Biology*, 3:0029, 2003.

P. Pfeiffer, O.S. Soyer, and S. Bonhoeffer. The evolution of connectivity in metabolic networks. *PLoS Biology*, 3(7):e228, 2005.

N.D. Price, J.L. Reed, J.A. Papin, S.J. Wiback, and B.O. Palsson. Network-based analysis of metabolic regulation in the human red blood cell. *Journal of Theoretical Biology*, 225:185–194, 2003.

T.A. Rapoport, R. Heinrich, and S.M Rapoport. The regulatory principles of glycolysis in erytbrocytes in vivo and in vitro. *Biochem. J.*, 154:449–469, 1976.

J. Raymond and D. Segré. The effect of oxygen on biochemical networks and the evolution of complex life. *Science*, 311:1764–1767, 2006.

V.N. Reddy, M.N. Liebman, and M.L. Mavrovouniotis. Qualitative analysis of biochemical reaction systems. *Comput. Biol. Med.*, 26:9–24, 1996.

I. Schomburg, O. Hofmann, C. Bänsch, A. Chang, and D. Schomburg. Enzyme data and metabolic information: Brenda, a resource for research in biology, biochemistry, and medicine. *Gene Funct. Dis.*, 3(4):109–18, 2000.

I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg. Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Research*, 32:D431–D433, 2004.

S. Schuster and C. Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.*, 2:165–182, 1994.

S. Schuster and C. Hilgetag. What information about the conserved-moiety structure of chemical reaction systems can be derived from their stoichiometry? *J. Phys. Chem.*, 99:8017–8023, 1995.

S. Schuster and T. Höfer. Determining all extreme semi-positive conservation relations in chemical reaction systems. a test criterion for conservativity. *J. Chem. Soc. Faraday Trans.*, 87:2561–2566, 1991.

S. Schuster, D.A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18:326–332, 2000.

S. H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

L. Stryer. *Biochemie*. Spektrum Akademischer Verlag, 2003. ISBN 978-3860253465.

Grace Development Team. Xmgrace. URL `http://plasma-gate.weizmann.ac.il/Grace/`.

A. Varma and B.O. Palsson. Metabolic flux balancing:basic concepts, scientific and practical use. *Bio/Technology*, 12:994–998, 1994.

A. Wagner and D. A. Fell. The small world inside large metabolic networks. *Proc. R. Soc. Lond. B*, 268:1803–1810, 2001.

D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

Z. Wunderlich and L.A. Mimy. Using the topology of metabolic networks to predict viability of mutant strains. *Biophysical Journal*, 91(6):2304–2311, 2006.

Danksagung

Als erstes möchte ich mich bei meinem Doktorvater Prof. Dr. Reinhart Heinrich für die großartige Unterstützung bei meiner Arbeit bedanken. Vorallem seine vielen Hinweise und Denkanstöße haben mir den Einstieg in die theoretische Biologie sehr erleichtert.

Natürlich möchte ich mich ganz besonders bei meiner Familie, insbesondere bei meiner Frau Irina, meinem Sohn Lennard und meinen Eltern für ihr Verständniss und ihre Geduld während dieser Arbeit bedanken.

Hervorheben möchte ich auch die angenehme Arbeitsatmosphäre in der theoretischen Biophysik, die viel zum Gelingen der vorliegenden Arbeit beigetragen hat. Wichtig waren auch die vielen Diskussionen und Anregungen, für die ich mich bei meinen Kollegen bedanken möchte.

Für die gute Zusammenarbeit, Diskussionen und Hinweise, sowie für das Korrekturlesen meiner Arbeit möchte ich insbesondere meinen Kollegen Oliver Ebenhöh, Nils Christian und Bernd Binder danken.

Die Arbeit wurde finanziell durch die Deutsche Forschungsgemeinschaft, im speziellen durch das Graduiertenkolleg "Dynamik und Evolution zellulärer und makromolekularer Prozesse" und den Sonderforschungsbereich 618 "Theoretische Biologie: Robustheit, Modularität und evolutionäres Design lebender Systeme" unterstützt.

Selbständigkeitserklärung

Ich versichere hiermit, die vorliegende Arbeit selbständig und ausschließlich unter Verwendung der angegebenen Mittel und ohne unerlaubte Hilfen angefertigt zu haben.

Berlin, den 19. November 2007

Thomas Handorf